Construction and assessment of tools for comparative structural genomic studies in white clover (*Trifolium repens* L.)

By

Melanie Febrer BSc.

A thesis submitted to the University of Wales in fulfilment of the Degree of Doctor of Philosophy

Department of Biological Sciences,
University of Wales Aberystwyth, Wales

Supervisors: Dan Milbourne PhD,
Glyn Jenkins PhD & Michael Abberton PhD.

November 2007

Acknowledgements

This work was completed with the help and encouragement of some very special people. I would like to express my enormous appreciation to them all.

First, to my supervisors Dr. Dan Milbourne, at Teagasc Crops Research Centre Carlow, Dr. Glyn Jenkins at the University of Wales, Aberystwyth and Dr. Michael Abberton at the Institute of Grassland and Environmental Research, Aberystwyth, for their support, guidance and patience throughout my research. It has been my privilege and pleasure to work with you three.

I would also like to thank Charlotte, Matthew, Andy, Rosemary and Tony from the Legume Breeding and Genetics Team at IGER for their assistance in my research. In particular I would like to thank the white clover breeder Terry Michaelson-Yeates, whose expertise and knowledge were precious from the start of this project.

I would like to thank my fellow researchers and friends, Timmy, Emma, Lucy, Steven, Paul, Andrew, Sarah, Olli, Killian and Mal for their help and kindness. I would also give a special thanks to Claire, Stephen, and Carlo for their helpful suggestions and incite during my research.

I am eternally indebted to my family, my Dad and Evelyne, my sisters Sophie, Estelle and Astrid and my aunt Fabienne, for their love, support and reassurance. To the Ryans for welcoming me in their family, for their patience and care.

Finally I would like to thank with all my heart, Robert who kept me upright and moving forward with his unconditional love and support. Without you, this thesis would not exist.

This thesis is dedicated to my Grandparents,

Marie and Jean Febrer,

For their unconditional love and belief.

DECLARATION

This work has not previously been accepted in substance of any degree and is not being concurrently submitted in candidature of any degree.

Signed…….

Date ……

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed.…….

Date……..

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for interlibrary loan, and for the title and summary to be made available to outside organisations.

Signed…….

Date…….

Abstract

Title: Construction and assessment of tools for comparative structural genomic studies in white clover (*Trifolium repens* L.)

Author: Melanie Febrer

White clover is an allotetraploid legume with a genome size of *ca* 960 Mb and is widely used in Irish grasslands to improve the nutritional value of the forage by increasing the relative amounts of nitrogen present. White clover is closely related to *Medicago truncatula,* the diploid (genome size of *ca* 500 Mb) model legume chosen to study the symbiotic genetic programmes of plants and facilitate the genetics and breeding of important legume crops.

This thesis describes the construction of a genetic map of a white clover mapping population composed of 94 F_1 progeny bred in the Institute of Grassland and Environmental Research (UK) from parents S1S4 and R3R4. The genetic map was constructed using amplified fragment length polymorphism markers and white clover microsatellite markers from previously published white clover maps. The thesis also details a glasshouse and field based phenotypic analysis carried out on morphological traits which allowed quantitative trait loci (QTL) to be detected on the genetic linkage map of white clover. Due to the high correlation between the traits measured, both conservation of QTLs across environments and co-localisation of QTLs for related traits were observed in this study.

A BAC library of one of the family parents, R3R4, was constructed using pIndigo BAC-5, consisting of over 37,000 clones with an average insert size of 85 Kb, representing a threefold genome coverage. A preliminary sequence analysis was carried out where 700 BAC clones were sequenced on both ends. Results showed that over 50% of the BAC-ends had a correspondence with the *Medicago* genome sequence; however only 16 of the BAC-end sequence pairs had homology within a span of 20 to 200 Kb in *M. truncatula.* The level of synteny between white clover and *M. truncatula* was assessed by comparative structural genomics, where five BAC clones of the white clover library were sequenced to sixfold coverage and their sequences were compared to the corresponding *M. truncatula* sequences.

Table of contents

List of figures

List of tables

Abbreviations

μg	Microgram
AFLP	Amplified Fragment Length Polymorphism
BAC	Bacterial artificial chromosome
BAC library	Bacterial Artificial Chromosomal library
BES	BAC-end sequence
cDNA	Complementary DNA
cM	centimorgan
DNA	Deoxyribo Nucleic Acid
EST	Expressed Sequence Tag
Kb	Kilo base pair
IM	Interval mapping
LG	Linkage group
LRT	Likelihood ratio test
Mb	Million base pair
MYA	Million years ago
N_2	Nitrogen
NH_3	Ammonia
O_2	Oxygen
PCR	Polymerase Chain Reaction
QTL	Quantitative Trait Locus
RAPD	Random Amplification of Polymorphic DNA
RFLP	Restriction Fragment Length Polymorphism
RILs	Recombinant inbred lines
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
SSR	Single Sequence Repeat
YAC	Yeast artificial chromosome

1.0 Introduction

1.1 Introduction

World agriculture includes two plant families of greatest importance, the cereals and grasses (Poaceae) and the legumes (Fabaceae) (Hymowitz, 1990). The Leguminosae or Fabaceae is one of the largest families of flowering plants with 18,000 species, which have been classified into 650 genera (Polhill, 1981). The Leguminosae is an extremely diverse family and constitutes one of humanity's most important groups of plants. The legumes are the most utilised plant family (Hymowitz, 1990), exploited for their chemicals, aesthetic value, timber, as cooking fuel, browse trees and shrubs, forage crops, pasture crops, cover crops, green manures, for feed and food, e.g. *Lupinus* (lupin), *Medicago sativa* (alfalfa), *Glycine max* (soybean) and *Trifolium* (clover).

Legumes are usually classified in terms of the time required to complete their life cycles. An annual legume like soybean germinates from seed, flowers, sets seed and dies within one growing season. In contrast, once established, perennial plants like alfalfa and white clover live for three or more years, and have potential to set seed each year. An intermediate group of legumes, biennials such as sweet clover, live for two years. These grow vegetatively the first year, and flower and die in the second year.

Of the three life cycle types of legumes, perennials are considered to be the most valuable for the environment, as they provide continuous ground cover, recycling of nutrients, and long-term carbon storage. The use of perennials also eliminates the need for annual reseeding (Sheaffer *et al.*, 2003).

Members of the legume family are characterised by their ability to fix nitrogen and the production of seeds and foliage that are usually rich in protein with a desirable amino acid composition. Many legumes are able to convert atmospheric nitrogen into nitrogenous compounds useful to plants. This is achieved by the presence of root nodules (which are visible to the naked eye) containing bacteria of the genus *Rhizobium*. These bacteria have a symbiotic relationship with legumes, fixing free nitrogen for the plants (Mylona *et al.*, 1995). In return legumes supply the bacteria with a source of fixed carbon produced by photosynthesis. This enables many legumes to survive and compete effectively in nitrogen poor conditions.

In terms of economic importance the Leguminosae are the most important family in the Dicotyledonae (Harborne, 1994). Legumes are second only to the grasses (cereals) in providing food crops for world agriculture. Major staple foods such as beans, soya, lentils, peas and chickpeas are all legumes.

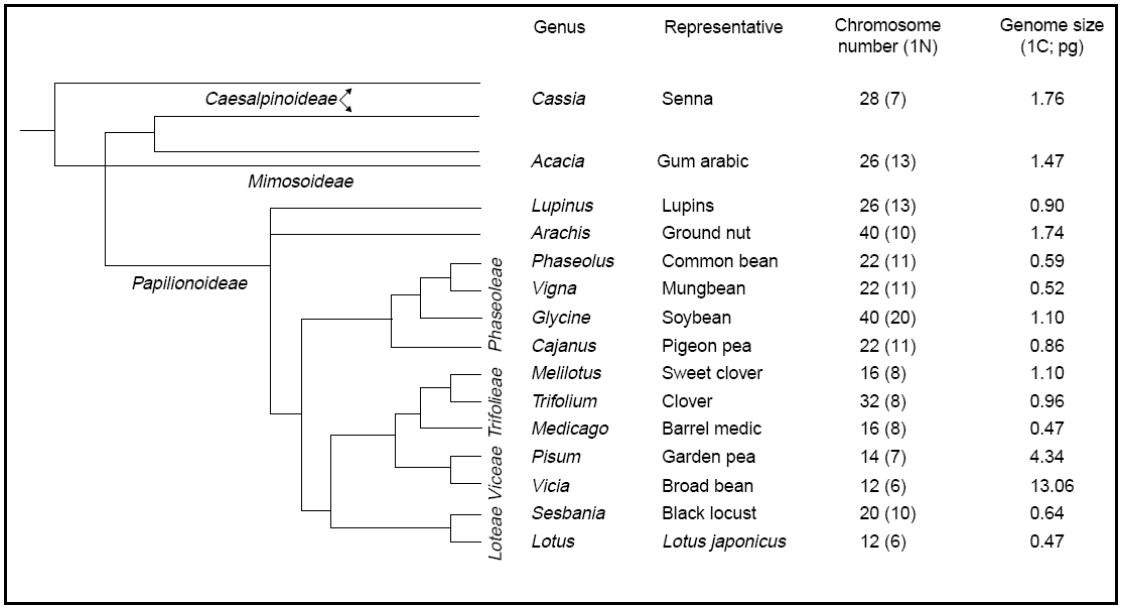
Forage legumes species are assigned different roles in grassland farming depending on their plant structure and features. Table 1.1 shows the traditional roles and characteristics of three major forage legume species. From this, one sees that low growing, stoloniferous, and competitive species such as white clover (*Trifolium repens* L.) are used as a component of grazed pasture whilst high yielding, and upright species such as alfalfa (*Medicago sativa* L.) are best suited to produce hay and silage in a monocropping system (Bouton, 1996). Another important role for forages is in soil composition. In these systems, forage grasses and legumes are used in watershed management and to reduce erosion runoff. This includes their use as cover crops and their role in conservation tillage systems (Bouton, 1996).

Table 1.1: Widely used forage legume species, their major role in forage production, and the characteristics that allow this role (Bouton, 1996).

Species	Role	Characteristics
Alfalfa, *Medicago sativa* L.	Hay and silage	Crown former, high yield, upright growth for easy harvesting, difficult to establish in grasses and non-tolerant of intensive grazing.
Red clover, *Trifolium pratense* L.	Hay and silage	Crown former, easy to establish in grasses, high yield, some tolerance to intensive grazing.
White clover, *Trifolium repens* L.	Pasture	Stoloniferous, easy to establish in grasses, competitive, tolerant to intensive grazing.

1.2 Papilionadeae subfamily

The legume family is usually divided into three subfamilies: Papilionoideae, Caesalpiniodeae and Mimosoideae (Bouton, 1996) (Figure 1.1). These subfamilies are sometimes recognised as three separate families: Papilionaceae, Caesalpiniaceae and Mimosaceae. The three subfamilies are generally identifiable by their flowers. The Mimosaceae and the majority of the Caesalpiniaceae are tropical or subtropical trees and shrubs.



Genus	Representative	Chromosome number (1N)	Genome size (1C; pg)
Cassia	Senna	28 (7)	1.76
Acacia	Gum arabic	26 (13)	1.47
Lupinus	Lupins	26 (13)	0.90
Arachis	Ground nut	40 (10)	1.74
Phaseolus	Common bean	22 (11)	0.59
Vigna	Mungbean	22 (11)	0.52
Glycine	Soybean	40 (20)	1.10
Cajanus	Pigeon pea	22 (11)	0.86
Melilotus	Sweet clover	16 (8)	1.10
Trifolium	Clover	32 (8)	0.96
Medicago	Barrel medic	16 (8)	0.47
Pisum	Garden pea	14 (7)	4.34
Vicia	Broad bean	12 (6)	13.06
Sesbania	Black locust	20 (10)	0.64
Lotus	*Lotus japonicus*	12 (6)	0.47

Figure 1.1: A phylogeny of legumes, featuring the three major subfamilies and details about selected crop species in the Papilionoideae. Estimates of the chromosome number and genome size are also shown (Young *et al.*, 2003).

The Papilionadeae, the largest of the three subfamilies, consists of two-thirds of all the genera and species of the legume family. It is also the most widespread, extending further into temperate regions than the other two subfamilies. The Papilionadeae contains most of the important crop species such as soybean (*Glycine max*), common pea (*Pisum sativum*), the medics (*Medicago sativa* and *M. truncatula*), forage crops like birdsfoot trefoil (*Lotus*) the clovers (*Trifolium repens* and *T. pratense*).

1.2.1 Soybean

The soybean (*Glycine max*) is a papilionoid legume and a member of the tribe Phaseoleae, subtribe Glycininae. The genus *Glycine* is unique within the subtribe with respect to several morphological and chromosomal characters. The soybean plant is a branched, non-frost tolerant, annual about one meter above ground level and two meters below ground level. The stem tissues are mostly primary, and the soybean flower is a standard papilionaceous flower with calyx of five united sepals.

Soybean is mainly self-pollinating enabling the easy production of separate breeding lines. Soybean is grown primarily for the production of seed, has a multitude of uses in the food and industrial sectors, and represents one of the major sources of edible vegetable oil and of proteins for livestock feed use. It is also used in various food

products, including tofu, soya sauce, and simulated milk and meat products. Soybean meal is used as a supplement in feed rations for livestock. Industrial use of soybeans ranges from the production of yeasts and antibodies to the manufacture of soaps and disinfectants.

1.2.2 Peas

The genus *Pisum* contains two species, *P. sativum* and *P. fulvum*, both with $2n = 14$ chromosomes. Cultivated peas are classified within *P. sativum* ssp. s*ativum,* which contains var. *sativum*, the horticultural types, and var. *arvense*, which are the fodder and winter types. The horticultural types are characterized by papilionaceous flowers that can be borne in singles or multiples on racemes that originate from the stem axes of viney upright plants. Flowers are usually white on horticultural types although some edible-podded cultivars have violet flowers. There is a vast range in pod types from small cylindrical pods to the large and flat edible pods. Similarly, there is tremendous variation in seed sizes, shapes and colours. The wrinkled seeded types are commonly used in the immature stage for freezing and canning while the smooth seeded types are used as dry peas. Peas are a cool season, annual crop planted in rotation with other processing crops such as potatoes, sweet corn, field corn, soybeans, and snap beans.

1.2.3 *Medicago* species

Medicago (medics) is a genus of about 50 species of annual and perennial herbs from Europe and Asia. Many species are cultivated as pasture legumes. The genus *Medicago* includes forage species of high symbiotic nitrogen fixation potential and high protein productivity. The members of the genus *Medicago* have flowers that grow usually several in short to fairly elongate spikelike clusters or heads on relatively short stalks from the leaf axils.

Alfalfa or lucerne (*Medicago sativa*) is the most important forage legume that has been cultivated for thousands of years. It is grown today on all continents except Antarctica. More than 33 million hectares are cultivated in the world. Cultivated types are outbreeding autotetraploids. Alfalfa is a perennial that is grown from seed. It is salt tolerant and also drought resistant. It makes excellent fodder (but is slightly toxic to many animals) and must be mixed with other forage. It is higher in protein

than either grasses or grain and oilseed crops, making it highly desirable for pasture and hay production, and is especially important in the dairy industry. Alfalfa's ability to fix atmospheric nitrogen makes it valuable for use in crop rotations, increasing the productivity of crops that follow it.

Medicago truncatula (commonly known as "barrel medic" because of the shape of its seed pods) is a forage legume commonly grown in Australia. It is an omni-Mediterranean species and closely related to the world's major forage legume, alfalfa. *M. truncatula* has a simple diploid genome (two sets of eight chromosomes) and can be self-pollinated. Section 1.7.2 on the use of a model legume will describe *M. truncatula* in more specific details.

1.2.4 Lotus

Lotus species are herbaceous perennial or annual legumes naturally distributed widely throughout the world and characterised by their leaves, which consist of five leaflets, three at the tip of the leaf stalk and two at the base where the leaf stalk joins the stem. The main agricultural species have many bright yellow pea-shaped flowers, which develop in late spring-early summer and are usually pollinated by bees. The flower heads become clusters of brown seedpods, which sometimes resemble a bird's foot.

Lotus is a large and extremely diverse genus consisting of approximately 150 species. The exact number is uncertain because there are several closely related groups that have very similar characteristics, making it difficult to separate them into individual species. They are also highly polymorphic, meaning that the same species can vary in appearance depending on the environmental conditions.

One perennial species, *Lotus japonicus*, is used as a model plant by legume researchers and is regarded as one of the most useful plants for legume study (VandenBosch & Stacey, 2003). It is valuable for research because it has a relatively small diploid genome (2n = 12). It also grows quickly and produces numerous small brown/blackish seeds. It is self-fertilising and can be regenerated from cell culture.

1.3 White clover (*Trifolium repens* L.)

White clover (*Trifolium repens* L.) has become a forage legume of interest in the need to reduce economic and environmental costs of livestock. Studies by Holliday (1989) and Young (1989) have shown that the use white clover based pasture is an economically viable option for sheep, beef and milk production. In Ireland, grazed grass/clover pastures cover approximately 80% of milk and 60% of beef production and the potential productivity of Irish grasslands is increased by the genetic improvement of component species through breeding of better varieties (especially white clover) (Connolly, 2001).

1.3.1 Taxonomy and distribution of white clover

White clover (*Trifolium repens* L.) is the most important true clover species for grazed swards within the genus *Trifolium* and the most important pasture legume in temperate zone. *Trifolium* belongs to the tribe Trifolieae of the subfamily Papilionoideae of the family Leguminosae (Australian Government, 2004). The genus consists of about 250 species distributed throughout the temperate and subtropical regions of the north and south hemispheres, particularly in Europe, northwest and central Asia, northeast Africa, parts of tropical Africa and South Africa, and western North and South America. Historically white clover was considered as a component of pastures in North-west Europe as far back as the seventeenth century.

1.3.2 Morphology of white clover

White clover is a potentially long-lived perennial, with multi-branched creeping stem called stolons. White clover plants develop an extensively branched tap root system from the primary seedling (Figure 1.2). Thus, adventitious roots systems with numerous lateral branches arise from the nodes that develop from the mother plant. The tap root system is however short lived. Roots of white clover grow to a similar depth to those of temperate grass species such as perennial ryegrass (*Lolium perenne*). White clover is a glabrous plant, with leaves that are normally trifoliate and ovate to circular. The first true foliage leaf is unifoliate and round, but later emerging leaves consist of three leaflets, with the edges either entire or serrate.

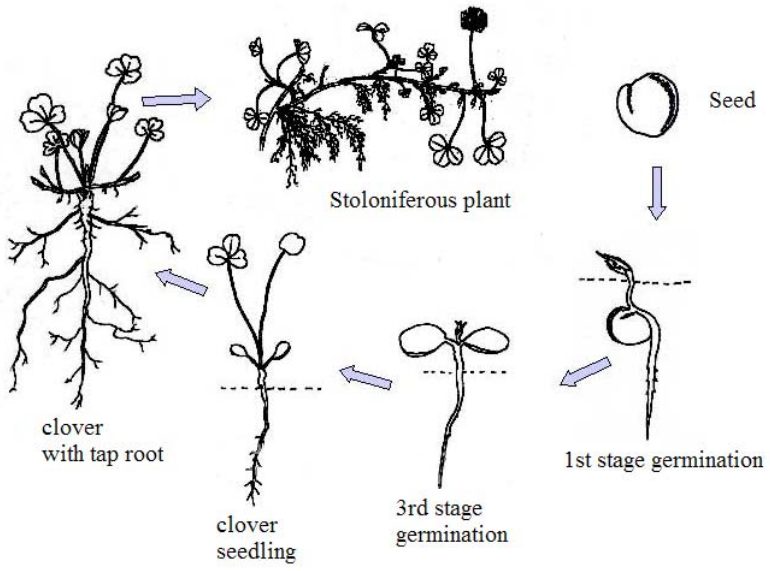


Figure 1.2. Stages of development of the white clover plant (http://www.iger.bbsrc.ac.uk/Practice/)

After the seedling stage, stolons emerge from the axils of the leaves to form a network of branched stolons. These stolons are organs, which accumulate carbohydrates in form of starch and their apices are the sites of leaf production.

The stolon is a key organ in the plant's survival and dry matter production of new leaves and branches it provides a means of perennation i.e. the ability to survive over winter and regenerate the following spring. The stolons are also a storehouse for reserves of carbohydrates and proteins necessary in survival and the regeneration of new leaves following defoliation, or winter.

Varieties of clover vary in their leaf and stolon characteristics, with big differences in such traits as leaf size, stolon thickness, profuseness of stolon production, content of reserve carbohydrates and protein in the stolons. Large differences also exist in the distance between the nodes on the stolon and hence the frequency of rooting along its length. All types of white clover are leafy, with short, almost prostrate stems. They bear white flowers in spherical clusters that are on stems slightly longer than the leaf petioles. New plants develop vegetatively on stolons near the soil surface.

1.3.3 Nitrogen fixation of white clover

Despite the importance of white clover nitrogen fixation to temperate agriculture, most research on the development and physiology of the legume nitrogen fixation symbiosis has been focused on other legumes. However, many of the important processes in infection, nodule development and nodule physiology probably vary only in detail among species. Inside the nodules, differentiated bacteria, called bacteroids, fix atmospheric nitrogen (i.e. reduce N_2 into NH_3) to the benefit of the plant.

Four distinct stages of infection of *Trifolium* species, including white clover, by *Rhizobium trifolii* have been described by (Crush, 1987) (Figure 1.3). They start with proliferation of rhizobia in the rhizosphere and root hair curling caused by bacterial secretion of two factors (a nucleic acid and a protein or polysaccharide) able to cause root hair deformation. These diffusible factors are apparently absent in non-invasive mutants. Initiation of the infection thread occurs by an invagination process and an open pore is formed by redirection of root hair wall growth. Growth of the infection thread remains extracellular within the root hair. Penetration of a polyploid root cortical cell by an infection thread stimulates the cell into meristematic activity and this, with nearby cells, constitutes the initial nodule primordium. White clover nodules are characteristically small (1.5 x 3 mm), pinkish-white, club-shaped to ellipsoidal, but occasionally are larger and fan-shaped. They have an apical meristem capable of enlarging the nodule as nodule cells degenerate progressively from the base. Individual polyploid cells of the nodule are infected by extensive proliferation of the infection thread, which releases bacteria into the cells. Here they multiply rapidly and change into the bacteroid form, each of which is surrounded in white clover by a membrane derived from the cell plasmalemma.

Nitrogen fixation consists essentially of the reduction of atmospheric dinitrogen gas to ammonia within the nodule, by action of the nitrogenase enzyme system. Nitrogenase consists of two components of different molecular weights (220,000 and 55,000 respectively), each containing Fe and the larger molecule also containing Mo. Neither component will alone fix nitrogen. The system requires oxygen, but nitrogenase is inactivated by O_2 at other than low partial pressures. Nitrogen fixation is confined to the bacteroids.

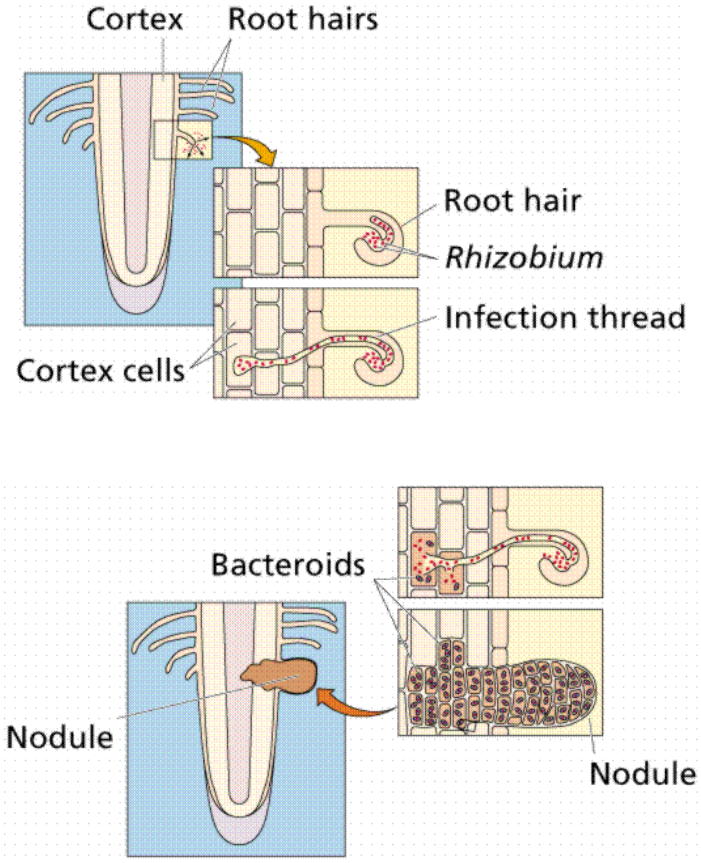


Figure 1.3. Stages of infection by *Rhizobium* bacteria.

1.3.4 Uses of white clover

White clover has been used in a mixed sward with grasses. It has been used for grazing, pasture hay and ground cover in horticultural situations. It is highly important in the dairy, meat and wool industries, significantly improving yields of these products (Australian Government, 2004). There are many advantages to using white clover in pastures. It has a high nutritive value because it supplies a rich source of proteins and minerals, and has high voluntary intake by grazing animals thereby making an important contribution to feed supply throughout the year. It is also adaptable to a wide range of soil and environmental conditions and combines well with many perennial grasses (Australian Government, 2004).

However the use of white clover can have some disadvantages. Bloat has been observed on white clover pastures and usually follows a heavy feeding or grazing period. Bloat is a digestive disorder characterized by an accumulation of gas in the first two compartments of a ruminant's stomach (the rumen and reticulum). White

clover is of particularly low persistence under nitrogen fertilization, but variability could also be influenced by species and cultivars of companion grasses and by interaction with the type and cultivar of white clover in grass/clover mixtures.

1.3.5 Genetics of white clover

White clover is a natural allotetraploid with 32 chromosomes (2n=4x=32). Allopolyploids originate from concurrent hybridisation and duplication of the genomes of different species and they essentially contain two or more diploid genomes. For example, they may contain two copies of genome A and two copies of genome B. In general, allopolyploids behave like diploids – each of the genomes behaves autonomously, although the multiple homoeologous genomes may be collinear. Due to divergence of the constituent genomes, and in some cased specific genetic mechanisms that maintain the allopolyploid state, inheritance in allopolyploids (which includes white clover) is disomic i.e. pairing and recombination of chromosomes during meiosis is restricted to homologous pairs (Australian Government, 2004). The predominantly outbreeding and disomic inheritance means that the white clover populations are composed of a heterogeneous mixture of highly heterozygous individuals. This results in high levels of genetic variation both within and between populations. The high genetic variability of white clover enables it to adapt to competitive microenvironments. This is an important attribute, as white clover does not naturally cross with other *Trifolium* species and therefore cannot gain genetic variation by forming hybrids (Australian Government, 2004). Recently, Ellison et al. (2006) studied the ancestral origin of white clover and identified two diploid *Trifolium* species, *T. occidentale* and *T. pallescens*, as putative ancestors for white clover.

1.3.6 Breeding of white clover

Plant breeding is mostly based on the selection of suitable genotypes possessing characteristics that determine improved adaptation and agronomic performance in the target environment (Evan, 1993). The objective of breeding many annual crops is to increase yield, while in forages, breeding is directed mainly towards enhanced tolerance of environmental constraints and improved total swards performances (Jahufer *et al.*, 2002). The method of breeding used in improving perennial forage species depends on the objective of the breeding effort and the species mode of

pollination (Table 1.2). Varieties derived from cross-pollinating species are advanced-generation populations derived from a selected group of parental lines, whereas self-pollinating varieties are homozygous pure lines developed from selected inbred lines (Bowley, 1997).

Table 1.2. Breeding methods for forage legumes (Bowley, 1997).

Mode of pollination	Breeding methods
Cross-pollinating	- Recurrent selection - Progeny testing - Backcrossing - Genetic variation
Self-pollinating	- Controlled hybridisation - Backcrossing - Genetic variation (mutation, transformation)

White clover is a cross-pollinating species and its breeding has been dominated by the classical approaches of mass selection and progeny testing, typical for and outbreeder species. Mass selection or phenotypic recurrent selection is a form of selection in which individuals are selected in each cycle based on their appearance or phenotype (Bowley, 1997). The method is simple to conduct and often provides a rapid response. There are two types of mass selection, Mass 1 and Mass 2, which differ in the type of pollen control used in each cycle of selection. Mass1 has one-parent control while Mass2 has two-parent control. In Mass1 selection, the selected plants are allowed to interpollinate with all plants in the population. Seeds are harvested from these selected female parents and used for the next cycle of selection. In Mass 2 selection, however, the pollen is restricted only to that of one group of plants selected. This enables selection for both male and female gametes during each cycle of recurrent selection.

Progeny testing is a breeding method in which parents are selected based on the performance of their progenies (Bowley, 1997). This method is preferred over phenotypic selection when the character under improvement had low heritability, that is, when there is large environmental influence on the phenotypic expression of the trait, as is the case for many of the traits of importance in clover breeding Evaluation of several related plants (i.e. progeny) tends to cancel the phenotypic deviations caused by environmental effects and enables determination of the relative genotypic value of potential parents. A better measure of each plant breeding value is obtained,

thus improving the selection gain. Two main types of progeny testing have been employed in the white clover breeding programme at Oak Park: (i) half-sibs, progenies that have one parent (usually the female) in common; (ii) full-sibs, progenies obtained from controlled crosses in which both parents are common (Figure 1.4). Following progeny testing, synthetic varieties are formed based on superior parents (in the half-sib testing process) or on seed from high performing full-sib families (in the full-sib testing process).

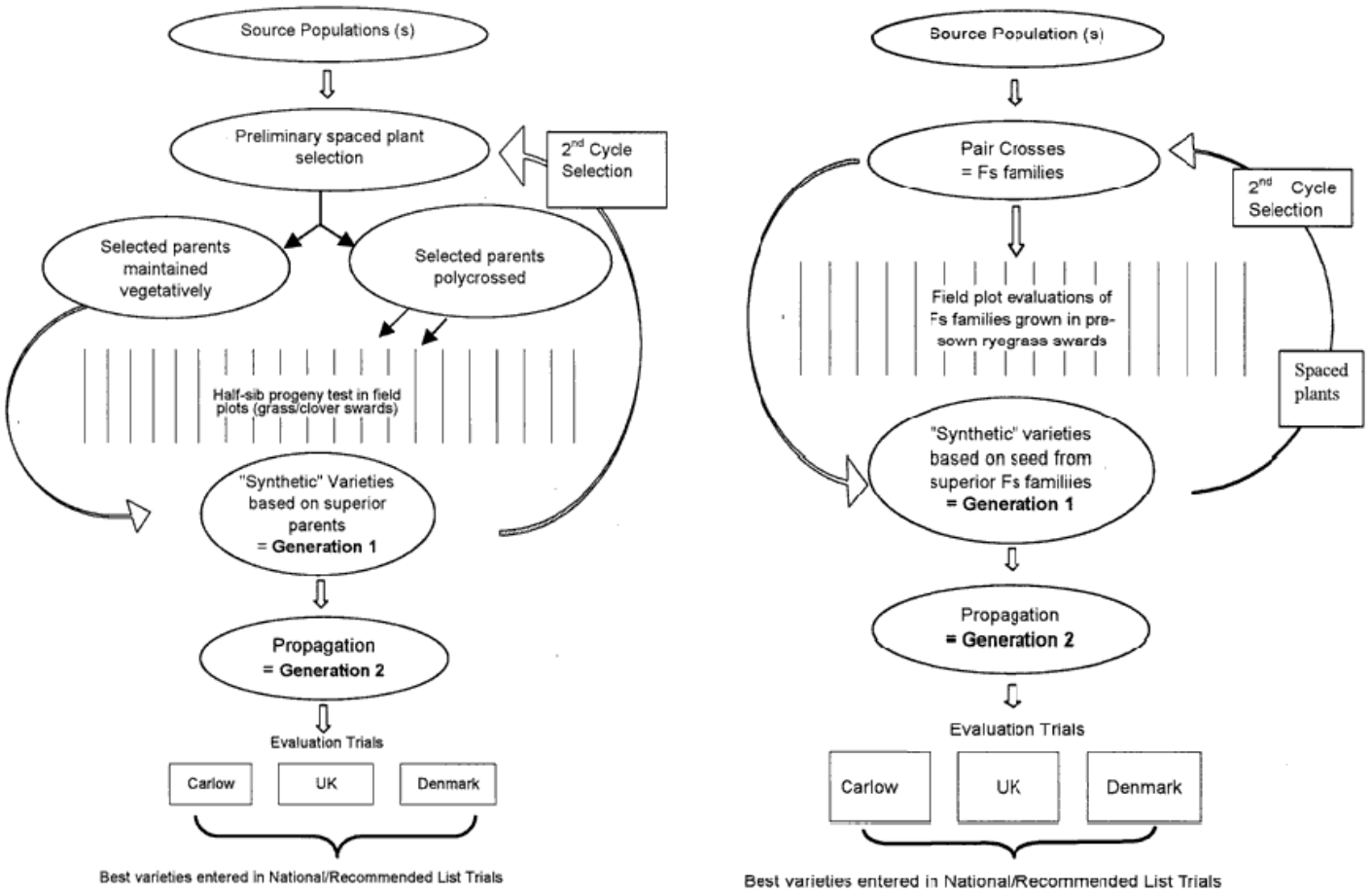


Figure 1.4. Two main types of progeny testing employed in the white clover breeding programme at Oak Park.

The primary objective in white clover breeding is to improve production from grazing animals. Important ingredients in the agronomic success of pasture legumes such as white clover include persistence, complementarity with associated grasses, quality for the grazing animals, compatibility with *Rhizobium* bacteria and compatibility with insect pollinators. The scope for improvement by the breeder is determined by the outcrossing nature of white clover and the consequent population structure, based on natural selection of heterozygous genetic combinations. Therefore, the breeder must retain heterozygosity by production of synthetic varieties or development of hybrids varieties (Williams, 1987). Along with perennial ryegrass, white clover was among the first species to be improved by scientific selection and breeding. Using the process described above, modern clover breeders have successfully produced new varieties of white clover that are persistent and reliable enough to ensure that these benefits are realised year after year. Major advances have been made allowing the commercialisation of cold hardy, grazing tolerant varieties that are more compatible with grasses. However, classical approaches to germplasm improvement could be enhanced with new molecular methods arising from molecular genetics. Molecular methods have the potential to reduce the duration of breeding programmes and also to increase the precision with which new trait combinations can be assembled. Marker assisted selection (MAS) is a complementary technology, for use in conjunction with more established conventional methods of genetic selection, for plant and animal improvement. In fact, for crops such as white clover, which require many years to estimate their performance in medium to long-term pastures, the use of MAS is more suitable for their use for annual crops.

Molecular dissection of complex traits and the development of molecular markers linked to genes and quantitative trait loci (QTL) controlling such traits may provide new tools for breeding, which can complement traditional breeding approaches (Newbury 2003). The following sections will review the technology and theory behind genetic linkage and the mapping of quantitative traits.

1.4 Genetic linkage mapping and molecular markers

The notion of genetic linkage was realised shortly after the rediscovery of Mendel's laws at the beginning of the century. Nevertheless, the development of comprehensive linkage maps for most species was not possible until the introduction of DNA based molecular marker technology. Advances in molecular biology techniques have provided the basis for uncovering virtually unlimited numbers of DNA markers. The utility of DNA-based markers is generally determined by the technology that is used to reveal DNA-based polymorphism (Bardakci, 2001). Molecular markers are relatively simple to detect, abundant throughout the genome even in highly bred cultivars, completely independent of environmental conditions and can be detected at virtually any stage of plant development. They have the big advantage that they are much more numerous than morphological markers, and they do not disturb the physiology of the organism (Jones *et al.*, 1997b). Molecular markers can be used for several different applications including: germplasm characterisation, genetic diagnostics, characterisation of transformants, and study of genome organisation, i.e. genetic linkage mapping and phylogenetic analysis. This section reviews the current status of DNA based marker technology, focusing on the most commonly used and powerful techniques currently employed.

DNA markers can be broadly classified into two categories depending on whether or not the potential exists to visualise all of the alleles present at the marker locus. Dominant markers generally allow the visualisation of only on allele per locus, while co-dominant markers generally have the potential to reveal all of the alleles at a locus.

The three majors types of dominant markers include Restriction Fragment Length Polymorphisms (RFLPs) and Simple Sequence Repeats (SSRs) or Microsatellites; while the two majors types of co-dominant markers include Amplified Fragment Length Polymorphisms (AFLPs) and Random Amplified Polymorphic DNA (RAPDs). The properties and relative advantages and disadvantages of these techniques are described in Table 1.3 (Rafalski & Tingey, 1993; Kumar, 1999).

Table 1.3. Comparison of the most common used molecular markers (Rafalski & Tingey, 1993; Kumar, 1999).

Dominance	**Co-dominant**			**Dominant**	
Feature	RFLPs	SSRs	SNPs	AFLPs	RAPDs
Principle	Endonuclease restriction Southern Blotting	PCR of simple sequence repeats	Direct sequencing Electronic SNPs	PCR amplification of genomic restricted fragments	DNA amplification with random primers
Type of polymorphism	Single base changes Insertions Deletions	Changes in length of repeats	Single base changes	Single base changes Insertions Deletions	Single base changes Insertions Deletions
Number of loci per assay	1.0-3.0	1.0-3.0	2.0	20-100	1.5-50
DNA required (μg)	10	0.05		0.5-1.0	0.02
PCR-based	No	Yes	Yes	Yes	Yes
Ease of use	Not easy	Easy	Very easy	Easy	Easy
Sequence information	No	Yes	Yes	No	No
Cost	Medium	High	Low	Medium	Low
Start-up cost	High	High	Low	High	Low

1.4.1 Co-dominant markers

1.4.1.1 Restriction Fragment Length Polymorphism (RFLP)

RFLPs were the first DNA markers to be used in 1980 by researchers in human genetics (Botstein *et al.*, 1980; Wyman & White, 1980). At this point, RFLPs occupied a major role in human genetic research and the first complete genetic linkage map of the human genome was based on RFLP markers (Doniskeller *et al.*, 1987).

RFLP markers are co-dominant, i.e. heterozygotes can be distinguished from homozygotes, thus providing a complete genetic picture at a single locus (Rafalski & Tingey, 1993). The amount of DNA needed for RFLP analysis is relatively large (5-10 μg), but a single Southern blot may be re-probed many times over a period of years, making this technology very efficient (Rafalski & Tingey, 1993).

RFLPs are molecular markers based on the differential hybridisation of cloned DNA-to-DNA fragments in a sample of restriction enzyme digested DNAs. This hybridisation-based technique requires the use of a library of DNA fragments cloned into some vector. These fragments may be from species under study or from related species. The library may be based on genomic or cDNA. RFLP does not require sequencing. Genomic DNA from the target organism is digested with one or more restriction endonucleases; the resulting fragments separated electrophoretically according to size, immobilised on a nylon membrane by Southern blotting and probed with DNA clones from the library. Fragments matching the probe DNA are visualised by autoradiography or the use of fluorescent labelling techniques.

There are a number of single gene traits that are frequently transferred from one genetic background to another by breeders. Genes conferring resistance to pathogens are a classic example. In tomato, RFLP markers have been identified that are tightly linked to genes for resistance to tobacco mosaic virus, *Fusarium* wilt, bacterial speck and root knot nematode (Tanksley *et al.*, 1989). In maize and lettuce respectively, researchers have established linkages between RFLP markers and genes for resistance to maize dwarf mosaic virus and downy mildew (Tanksley *et al.*, 1989).

1.4.1.2 Single Sequence Repeats (SSR)

Single Sequence Repeats (SSRs) also known as microsatellites are stretches of DNA, consisting of tandemly repeating mono-, di-, tri-, tetra- or penta-nucleotide units, that are arranged throughout the genomes of most eukaryotic species (Powell *et al.*, 1996; Dograr & Akkaya, 2001). The existence of dinucleotide repeats was first acknowledged in 1982 (Hamada *et al.*, 1982). Subsequent studies have confirmed both the abundance and ubiquity of microsatellites in eukaryotes (Tautz & Renz, 1984). The number of repetitive motifs is very variable not only between species, but also between closely related individuals (Miyao *et al.*, 1996). Because SSR polymorphism often exhibits co-dominant Mendelian inheritance, SSRs are useful for genetic mapping (Miyao *et al.*, 1996). Several authors have reported that the abundance of SSR motifs differs between plants and vertebrates (Lagercrantz *et al.*, 1993); specifically, the d(AC/GT)n motif is abundant in humans, and is infrequently observed in plant genomes. However other SSR motifs are also useful for detecting polymorphism and have been used in many approaches to genetic mapping, e.g. those for maize (Senior & Heun, 1993), rice (Wu & Tanksley, 1993), *Arabidopsis* (Bell & Ecker, 1994), soybean (Morgante *et al.*, 1994), barley (Becker & Heun, 1995) and tomato (Broun & Tanksley, 1996).

Polymorphism is revealed by PCR-amplification from total genomic DNA, using two unique primers, composed of short lengths of nucleotides that flank and hence define the microsatellite locus. Amplification products obtained from different individuals can be resolved electrophoretically to reveal polymorphism (Powell *et al.*, 1996). The uniqueness and value of microsatellites arises from their multiallelic nature, co-dominant transmission, ease of detection by PCR, relative abundance, extensive genome coverage and requirement for only a small amount of starting DNA (Table 1.3).

The first application of microsatellites in plants has been in cultivar identification, where microsatellites have been used to genotype such diverse material as soybean (Powell *et al.*, 1996). Microsatellites can be used in genetic fingerprinting, parental analysis, genome mapping and marker-directed plant breeding (Broten, 2000). Marker-directed plant breeding uses the detection of microsatellites to signal specific

traits in the plant. This helps to speed plant breeding, as analysis of the markers lets the plant breeders recognize plant traits sooner (Broten, 2000).

1.4.1.3 Single nucleotide polymorphism (SNP)

Contrary to RFLP and SSR markers which need gel-based assays, a new emphasis is now shifting towards the development of molecular markers, which can be detected through non gel-based assays. One of the most popular of these non gel-based marker systems is SNP, which represents sites, where DNA sequence differs by a single base (Gupta *et al.,* 2001). SNPs, at a particular site in a DNA molecule should in principle involve four possible nucleotides, but in actual practice only two of these four possibilities have been observed at a specific site in a population. Consequently, SNPs are biallelic as against the polyallelic nature of the once much preferred SSRs. This polymorphism has been shown to be the most abundant, so that at least one million SNPs should be available, only in the non-repetitive transcribed region of the human genome (Collins *et al.,* 1999).

The reason for increasing popularity of SNPs as simple bi-allelic co-dominant markers is the recent need for very high densities of genetic markers for the studies of diseases, and the recent progress in polymorphism detection and genotyping techniques (Vignal *et al.,* 2002). Several different routes to the discovery of SNPs may be taken. The choice of a method for a particular assay depends on many factors, including cost, throughput, equipment needed, difficulty of assay development, and potential for multiplexing.

Although numerous approaches for SNP discovery have been described, including some also currently used for genotyping, the main ones are based on the comparison of locus-specific sequences, generated from different chromosomes. The simplest, when targeting a defined region for instance containing candidate genes, is to perform direct sequencing of genomic PCR products obtained in different individuals. Depending on the frequency of polymorphisms in the germplasm under investigation, it may be beneficial to pre-screen amplicons for the presence of polymorphisms. Several methods may be employed, including denaturing high-pressure liquid chromatography, single-strand conformational polymorphism (SSCP), or one of several chemical or enzymatic cleavage methods (Rafalski, 2002).

1.4.2 Dominant markers

1.4.2.1 Random Amplification of Polymorphic DNA (RAPD)

Even though the RFLP assay has been the choice for many species to measure genetic diversity and construct genetic linkage maps, this technique is in general time consuming and laborious. Polymerase chain reaction (PCR) technology has become a widespread research technique and has led to the development of several novel genetic assays based on selective amplification of DNA. The discovery that PCR with random primers can be used to amplify a set of randomly distributed loci in any genome facilitated the development of these genetic markers. The RAPD technique was first employed by Williams *et al.* (1990) to examine human DNA samples from anonymous individuals. RAPDs have typically been used for genetic mapping and studies of population genetic structures (Aaggaard *et al.*, 1998).

The RAPD technique is performed on a genomic DNA template and primed by an arbitrary oligonucleotide primer, resulting in the amplification of several discrete DNA products (Rafalski & Tingey, 1993; Jones *et al.*, 1997a). Each product is derived from a region of the genome that contains two short segments in inverted orientation, on opposite strands, that are complementary to the primer and sufficiently close together for the amplification to work (Jones *et al.*, 1997a). This PCR-based technique requires neither cloning nor sequencing of DNA. It can detect several loci simultaneously.

Although the RADP method is relatively fast, cheap and easy to perform, the issue of reproducibility has been of much concern. This PCR-based fingerprinting method has the major disadvantage that it is very sensitive to the reaction conditions, DNA quality and PCR temperature profiles, which limit its applications.

1.4.2.2 Amplified Fragment Length Polymorphism (AFLP)

The AFLP technique, developed in 1993 (Zabeau & Vos, 1993; Vos *et al.*, 1995), is a technique for fingerprinting genomic DNA. It is based on the detection of genomic restriction fragments by PCR amplification and can be used for DNAs of all origin and complexity. DNA is cut with restriction enzymes, and double-stranded adapters are ligated to the ends of the DNA-fragments to generate template DNA for amplification. The sequence of the adapter and the adjacent restriction site serve as

primer binding sites for subsequent amplification of the restriction fragments. Selective nucleotides are included at the 3' ends of the PCR primers, which therefore can only prime DNA synthesis from a subset of the restriction sites. Only restriction fragments in which the nucleotides flanking the restriction site are complementary to the selective nucleotides will be amplified (Figure 1.5).

The restriction fragments for amplification are generated by two restriction enzymes, a rare cutter and a frequent cutter. The AFLP procedure results in predominant amplification of those restriction fragments, which have a rare cutter sequence on one end and a frequent cutter sequence on the other end. The rationale for using two restriction enzymes is as follows. (i) The frequent cutter will generate small DNA fragments, which will amplify well and are in the optimal size range for separation on denaturing gels. (ii) The number of fragments to be amplified is reduced by using the rare cutter, since the rare cutter/frequent cutter fragments are amplified. This limits the number of selective nucleotide needed for selective amplification. (iii) The use of two restriction enzymes makes it possible to label one strand of the ds PCR products, which prevents the occurrence of 'doublets' on the gel due to unequal mobility of the two strands of the amplified fragments. (iv) Large numbers of different fingerprints can be generated by the various combinations using a low number of primers, which can be used in many organisms.

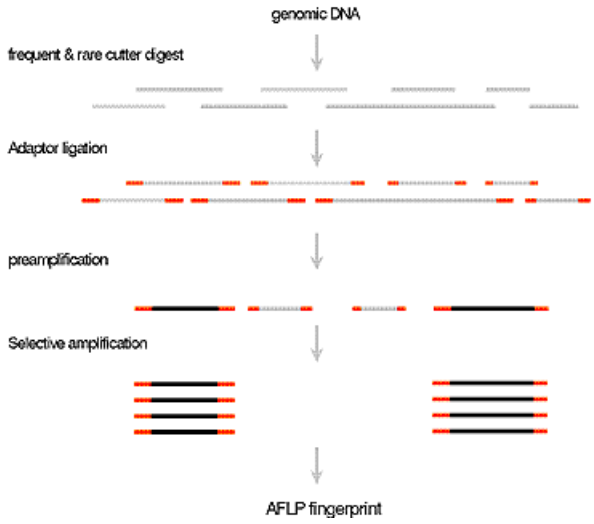


Figure 1.5. An overview of the AFLP technology.

AFLP markers advantageous in that they have:

- Taxonomic scope: AFLP markers can be generated for any organism with DNA, and no prior knowledge about the genomic makeup of the organism is needed.
- Error levels: AFLP amplifications are performed under conditions of high selectivity (at high stringency), thus eliminating the artifactual variation that is seen routinely in RAPD.
- Quantity of DNA: AFLP analysis requires minimal amounts of DNA (Table 1.3) and partially degraded samples can be used.
- Time efficiency: AFLP markers can be generated at great speed.
- Mendelian inheritance: AFLP markers segregate in a Mendelian fashion and can be used for population genetics and QTL (Quantitative Trait Loci) analysis.
- Resolution: Because of the nearly unlimited number of markers that can be generated with AFLP, using a series of different primer combinations, at least some AFLP markers will be located in variable regions and thus reveal even minor genetic differences within any given group of organisms.

AFLP methods rapidly generate hundreds of highly reproducible markers from DNA of any organism; thus, they allow high-resolution genotyping and fingerprinting quality (Mueller & Wolfenbarger, 1999). The key feature of AFLP is its capacity for the simultaneous screening of many different DNA regions distributed randomly throughout the genome. AFLP markers have been applied to evaluate gene flow and dispersal, outcrossing, introgression and cases of hybridisation. The high resolution of AFLP markers also enables testing for clonal identity between individuals and thus permits interferences about sexual versus asexual modes of reproduction (Mueller & Wolfenbarger, 1999).

1.5 Linkage mapping with molecular markers

Until recently, most molecular marker based linkage analyses have been performed in populations derived from the F_1 of crosses between two inbred (homozygous) diploid parents (eg, backcrosses, F_2, recombinant inbred lines and doubled haploids). This is due to the fact that many agronomically important crop species are fully self

fertile, and that in such populations linkage analysis is simplified because only two alleles segregate per locus and the linkage phase (coupling or repulsion) of all markers is known.

Individual population types have different advantages for mapping. F_2 populations, when used with co-dominant markers, allow recombination events on both homologous chromosomes to be followed (Allard, 1956; Reiter *et al.*, 1992). Backcross populations allow the allelic contribution of one parent (the recurrent) to be fixed at all loci, allowing a map to be produced using markers segregating from the other (nonrecurrent) parent. Recombinant inbred lines (RILs) and doubled haploid populations provide immortal populations of homozygous individuals which are ideal for the construction of genetic linkage maps using dominant markers (Burr *et al.*, 1988; Powell *et al.*, 1992; Reiter *et al.*, 1992). In addition, RILs offer an increase in map resolution, since they are the product of several rounds of meiosis (Burr *et al.*, 1988).

Linkage analysis is more complicated in populations derived from the progeny of crosses between non-inbred parents of an outbreeding species. Markers vary in the number of segregating alleles, one or both parents may be heterozygous at a locus, and frequently, the linkage phase of the markers is unknown (Maliepaard *et al.*, 1998). White clover is an outbreeder, highly heterozygous species, and generally self incompatible, precluding the possibility of obtaining inbred lines. The construction of genetic linkage maps in crosses between heterozygous parents has been presented by Ritter *et al.* (1990). In addition, many computer packages (such as JoinMap, which is used in this study) used to construct genetic linkage maps can now deal with segregation data obtained from crosses between heterozygous parents for both dominant and co-dominant markers (Van Ooijen & Voorrips, 2001).

Genetic linkage maps are essential for genome analysis, including map-based cloning and construction of physical maps (Hayashi *et al.*, 2001). A genetic linkage map helps to understand the structure, function and evolution of the genome. It can be an important tool for agricultural crop improvement. Recent work has shown that the genetic maps of many closely related species are quite similar with respect to the content and location of genes.

Several molecular marker based linkage maps are now available in the *Trifolium* species. A restriction fragment length polymorphism (RFLP) linkage map of diploid red clover (*T. pratense*) has been developed in a backcross population (Isobe *et al.*, 2003). The map contains 157 RFLP markers and one morphological marker on seven linkage groups. The total map distance was 535.7 cM and the average distance between two markers was 3.4 cM. A genetic linkage map of red clover (*T. pratense*) was constructed in a F_1 mapping population (HR x R130) (Sato *et al.*, 2005). The map consisted of 1305 microsatellite markers as well as the previously developed 167 restriction fragment length polymorphism markers A total of 1434 loci detected by 1399 markers were successfully mapped onto seven linkage groups totaling 868.7 cM in length. A third molecular linkage map of red clover has been developed from a two way test-crossed population (Herrmann *et al.*, 2006). The map contained 216 AFLP markers and 42 SSR markers, resulting in a total map length of 444.2 cM.

The first molecular marker-based genetic map of white clover was developed by Jones *et al.* (2003). This map was a framework genetic map of white clover constructed using an F_2 progeny set derived from the intercross of fourth and fifth generation inbred genotypes carrying a self-fertile mutation. White clover SSR (TRSSR) and AFLP markers were used to derive a map with 135 markers assigned to 18 linkage groups (LGs) with a total map length of 825 cm. In 2004, Barrett *et al.* constructed a comprehensive genetic linkage map of white clover using 92 F1 progeny from a pair cross between two highly heterozygous genotypes 364/7 and 6525/5. This map was constructed using the combination of in silico white clover EST-SSRs and genomic SSRs. Map length was 1,144 cM and spanned all 16 homologues. A third map of white clover was constructed using AFLP markers, white clover genomic SSRs and EST-SSR developed from *Medicago truncatula* (Jones, 2005). This map was developed using a white clover family derived from a cross between parents representing the end points of divergent selection for stolon characteristics. The length of this low-density framework map was of 359 cM.

Even though mapping in white clover has been widely studied, limited marker transfer has occurred to align the maps, and the number of white clover genetic markers accessible in the public domain is still relatively small. This year, Zhang *et al.* were the first to attempt the alignment of previously published maps of white

clover with molecular markers from other legume species (Zhang *et al.*, 2007). An F_1 population from a cross between two highly heterozygous genotypes, was used for genetic mapping. The map consists of 343 SSR primer pairs from various species including white clover, red clover, *M. truncatula*, and soybean. Linkage groups for all eight homoeologous chromosome pairs of allotetraploid white clover were detected and map length was estimated at 1,877 cM.

1.6 Mapping quantitative trait loci (QTLs) in white clover

The accessibility of dense molecular marker based maps has improved the mapping of a range of monogenic traits. Mapping these traits is simplified by the fact that they generally exhibit classical Mendelian inheritance patterns, and can be submitted for analysis in the same dataset as the molecular marker segregation data. Quantitatively inherited traits are those that have a strong genetic component, but which cannot be shown to be controlled by individual loci due to a lack of discrete phenotypic classes. The theory of QTL mapping was first described in 1923 by Sax, where he noted that seed size in bean (a complex trait) was associated with seed coat color (a simple, monogenic trait) (Sax, 1923). It was then suggested that if the segregation of simply inherited monogenes could be used to detect linked QTLs, and then it should eventually be possible to map and characterise all the QTLs involved in complex traits (Thoday, 1961). Modern QTL mapping is essentially the fulfillment of this idea, with the key innovation being that defined sequences of DNA act as the linked monogenic markers. With the development of comprehensive DNA marker maps, it is now possible to search for QTLs throughout the genomes of most crop species. This has had the profound effect of moving the focus in studies of polygenic traits to questions about the chromosomal locations, gene actions, and biological roles of specific loci involved in complex phenotypes (Tanksley, 1993).

In its simplest form, QTL mapping involves testing DNA markers throughout a genome for the likelihood they are associated with a region of the genome influencing a quantitative trait (probably because that region contains one or more genes influencing the trait). Individuals in a suitable mapping population are analysed in terms of DNA marker genotypes and the phenotype of interest. For each DNA marker, the individuals are split into classes according to marker genotype. Mean and variance parameters are calculated and compared among the classes. A

significant difference between classes suggests there is a relationship between the DNA and the trait of interest – in other words, the DNA marker is probably linked to a QTL (Young, 1996). More advanced techniques, such as interval mapping (Lander and Botstein, 1989) have been developed, which used one marker interval at a time to construct a putative QTL for testing by performing a likelihood ratio test (LRT) at every position in the interval. With a fine-scale genetic map throughout the genome, interval mapping can be performed at any position covered by markers to produce a continuous LRT statistical profile along chromosomes. The position with the significantly largest LRT statistic in a chromosome region is an estimate of QTL position (Kao *et al*., 1999).

To date, only few QTL mapping studies have been performed on the morphogenetic and reproductive development traits in white clover (Jones, 2005; Cogan *et al.*, 2006). Cogan *et al*. (2006) analysed a range of vegetative morphogenesis and reproductive morphogenesis and development traits in both individual and multi-environment combined analyses. This QTL mapping study, based on the genetic map from Jones *et al*. (2003), detected QTL for the majority of traits and found that and the locations and magnitudes of QTL effects were compared between individual and combined analyses. In another study, QTL mapping was performed on a range of morphological traits in a population derived from a cross between parents representing the end points of divergent selection of stolon characteristics (Jones, 2005). Large effect QTLs were identified for stolon length (50%) and leaf width (60%); however this analysis was limited by the low number of mapped markers (Jones, 2005).

Other QTL mapping studies have been focused on QTL for seed production (Barrett *et al.*, 2005), for root-knot nematode resistance (Barrett *et al*., 2005b). Following the development of a comprehensive genetic linkage map of white clover, Barrett et al. used the resource to identify QTLs that regulate seed yield and three of its components, inflorescence density, yield per inflorescence and thousand-seed weight (Barrett *et al.,* 2005a). A total of 23 QTLs related to seed production were identified, and both parents harboured valuable alleles for seed yield and the three component traits. The other study by Barrett *et al,.* (2005b) used bulk-segregant analysis (Michelmore *et al.,* 1991) to determine the genetic mapping of a root-knot nematode

resistance locus in *Trifolium*. The analysis was carried out on a pair cross between a clover root-knot nematode (CRKN) -resistant and CRKN-susceptible genotype of *T. semipilosum* cultivar 'Safari'. The two parents, TsR, TsS, and bulked resistant and bulked susceptible F_1 progeny were screened using white clover SSR markers from Barrett *et al.* (2004).

Recently, QTL analysis for seed yield components was performed on red clover (*T. pratense*) (Herrmann *et al.*, 2006). A total of 38 QTLs were identified for eight seed yield components and QTLs for several of these traits were often detected in the same genome region. Two genome regions containing four or five QTLs for different seed yield components, respectively, were also identified representing candidate regions for further characterisation of QTLs.

1.7 The use of model species

Over the last few years, research has been focused on developing model species, e.g. *Arabidopsis thaliana* for dicotyledons (Meinke *et al.*, 1998b) and two model legumes, *Lotus japonicus* (Pedrosa et al., 2002) and *Medicago truncatula* (Huguet & Prosperi, 1996; Frugoli & Harris, 2001). The main advantage of such model species is their relatively small genome size (Table 1.4), inbreeding mode of reproduction and rapid generation time, which make them much more tractable for the study of key processes in plants in a laboratory environment. Thus, for the study of other plants with larger genomes the knowledge gained from model species will be of major help.

Table 1.4. Comparison of genome sizes of the different model species.

Species	*Arabidopsis thaliana*	*Lotus japonicus*	*Medicago truncatula*
Family	Dicotyledons	Legumes	Legumes
Genome size	125 Mb	470 Mb	500 Mb

1.7.1 Models for flowering plants – *Arabidopsis thaliana*

In the 1980s there was a growing awareness that significant investments in studies of many different plants, such as corn, oilseed rape, and soybean. Researchers soon realised that understanding the physical make-up and development of plants is most quickly achieved by studying one specific plant, which has similar characteristics or

is related to important plant species (Friesen, 1998). This could be accomplished by turning to a model species that many scientists then study.

With this in mind, biologists began to search for a model organism suitable for detailed analysis using the combined tool of genetics and molecular biology (Meinke *et al.*, 1998b). During the last 8 to 10 years, *Arabidopsis thaliana* has become universally recognised as a model plant for such studies. *Arabidopsis thaliana* is a member of the mustard family (Cruciferae or Brassicaceae) with a broad natural distribution throughout Europe, Asia and North America (Meinke *et al.*, 1998b). Several features of *A. thaliana* are favoured by plant researchers. It develops, reproduces and responds to stress and disease in much the same way as many crop plants; it produces many seeds and is easy and cheap to grow; compared to other plants, it has a small genome (125 Mb) organised into five chromosomes and containing an estimated 20,000 genes (Meinke *et al.*, 1998b). Therefore, it was the first plant to have its genome sequenced due to an international coordinated program (Arabidopsis Genome Initiative, 2000).

Information from *Arabidopsis* can be used in many different ways. For example, flowering development has received special attention and many genes involved in this process have now been characterised at the genetic and molecular levels (Weigel, 1995). Scientists have found the *Arabidopsis* genes which promote early flowering such as the *PHYTOCHROME B* gene (Halliday *et al.*, 1994), the *SPINDLY* gene (Huala & Sussex, 1992) and the *TERMINAL FLOWER* gene (Alvarez *et al.*, 1992), and those which delay flowering such as the late flowering *CONSTANS* gene (Redei, 1962; Putterill *et al.*, 1995), the *FRIGIDA* gene (Clarke & Dean, 1994) and the *GIGANTEA* gene (Redei, 1962; Araki & Komeda, 1993). The discovery of these genes has permitted detection and manipulation of their orthologues in crop plants, which provided opportunities for breeders to develop improved varieties, depending on whether they required early or late flowering periods. Moreover, discovery of these genes has given insights into the molecular basis of quantitative traits; for instance, two major QTLs identified in *Brassica rapa* and *B. napus* corresponded to late flowering genes *CONSTANS* and *FRIGIDA* in *Arabidopsis* (Osborn *et al.*, 1997; Axelsson *et al.*, 2001) However, *Arabidopsis* is not a universal model because this plant does not encompass all of the diverse physiological, developmental, and

environmental processes seen throughout the plant kingdom (VandenBosch & Stacey, 2003).

1.7.2 Models for legume plants – *Medicago truncatula*

In recent years, investigators have sought a legume species that could serve as a model genetic system for certain developmental problems that cannot be studied in *Arabidopsis* (Cook *et al.*, 1997b). In order to understand the genetic system of legume species and to facilitate isolation and characterisation of genes responsible for legume-specific phenomena, including plant-microbe interactions and symbiotic nitrogen fixation (Nakamura *et al.*, 2002a), two model legumes, *Medicago truncatula* and *Lotus japonicus*, have been developed independently. Key attributes of these species include their diploid, autogamous nature, short generation times, and genome sizes only three to four times that of *Arabidopsis* (Cook *et al.*, 1997b). In contrast with the small, diploid genomes of the model legumes, the genomes of the important crop legumes are complex and often contain large quantities of repetitive DNA (Harrison, 2000). Most important European legume crops such as pea (*Pisum sativum*), faba bean (*Vicia faba*), chickpea (*Cicer arietinum*), lucerne (*Medicago sativa*) and clovers (*Trifolium* sp.) are members of the same phylogenetic group of legumes, the Galegoid, and the model plant *M. truncatula* is also in this group (Young *et al.*, 2003). The other model species, *L. japonicus*, is more distantly related and so *M. truncatula* is the most appropriate model for the most significant European legume crops. However, the development of these two different models can be justified because they exhibit two developmental systems for nodulation as well as other differences (VandenBosch & Stacey, 2003). *L. japonicus* forms determinate nodules in which the root subepidermal cortical cells initiate nodule formation and a persistent terminal nodule meristem does not develop. In contrast, *M. truncatula* nodules initiate from the division of inner cortical cells and continue to grow from a terminal, persistent meristem (VandenBosch & Stacey, 2003).

Lotus japonicus is a typical model legume with the following characteristics: short life cycle (2-3 months), self-fertility, diploidy (n=6), small genome size (472.1 Mb) (Sato *et al.*, 2001) and is readily transformable by *Agrobacterium*. *M. truncatula* is one of the 32 described *Medicago* species that are annual, autogamous diploids

(Cook *et al.*, 1997b). It has a simple diploid genome (2n=2x=16; 500-550 Mb) and can be self-pollinated, greatly facilitating genetic analysis.

Over the past decade several research groups have developed the tools and infrastructure necessary for basic research in *M. truncatula*, including efficient transformation systems, high-throughput systems for forward and reverse genetics including insertional mutagenesis (Tadege *et al.*, 2005), RNA interference (Limpens *et al.*, 2004), gene TILLING, well-characterized cytogenetics, 1,996 BACs sequenced (ftp://ftp.tigr.org/pub/data/m_truncatula/) plus detailed physical and genetic maps (mtgenome.ucdavis.edu/), gene knockout systems involving T-DNA and Tnt1 (Scholte *et al.*, 2002; d'Erfurth *et al.*, 2006), 226,923 *M. truncatula* expressed sequence tags (ESTs) (http://www.*Medicago*.org/genome/downloads/Mt1/ and http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=*Medicago*), with corresponding microarray and DNA chips and a collaborative research network (http://*Medicago*.org). Research efforts on *M. truncatula* encompass a broad range of issues in plant biology, from studies of population biology and resistance genes, to the molecular basis of symbiotic interactions and micronutrient homeostasis. Combined with the rapidly emerging sequence of its genespace, *M. truncatula* provides an impressive array of genomic tools to legume biologists. Because the *M. truncatula* sequencing effort is clone-by-clone, syntenic comparisons between these two genomes and with other plant taxa will be straightforward and highly informative.

1.8 Genome sequencing

The elucidation of the complete sequence of the genomes of many model organisms is probably the most important scientific accomplishment in the last decade of the 20th century and the beginning of the 21st century. Essentially, the genome sequence provides an insight view of the information needed for understanding the biology of model organisms.

There have been many debates on how to access the essential sequence information of large plant and animal genomes. There are essentially two ways to sequence a genome (Figure 1.6).

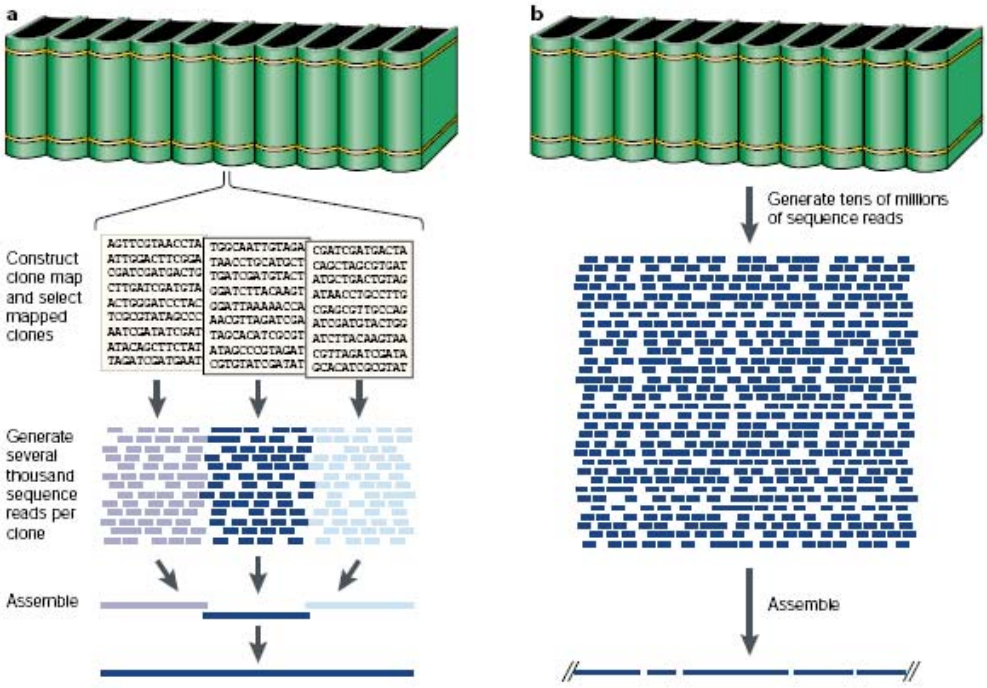


Figure 1.6. The two main ways to sequence a genome: a. Schematic overview of clone-by-clone shotgun sequencing. b**.** Schematic overview of whole-genome shotgun sequencing. (Green, 2001).

1.8.1 Clone-by-clone sequencing

The most frequently used method for establishing the sequence of large, complex genomes involves the shotgun sequencing of individual mapped genomic clones, also referred to as a clone-by-clone sequencing (Figure 1.6). The process of clone-by-clone shotgun sequencing can be conceptually divided into a series of discrete and sequential steps.

This method first requires the construction of libraries consisting of large DNA inserts, i.e. 50 Kb or greater, such as yeast artificial chromosomes (YACs), P1 phage vectors and bacterial artificial chromosomes (BACs). Large insert DNA libraries are necessary for construction of overlapping contigs of complex genomes as they provide long-range continuity for the genome and further enable the physical and genetic maps to be aligned (Klein *et al.*, 1998). Increasingly, plant scientists are focusing on BAC libraries because they are relatively simple to construct, inserts are easy to recover, the frequency of rearrangement is low (Danesh *et al.*, 1998) and they are capable of stably maintaining large inserts of genomic DNA (Ding *et al.*, 1999). Efficient approaches for constructing BAC-based physical maps have been developed in recent years (Mozo *et al.*, 1999; Chen *et al.*, 2002). Briefly, restriction

enzyme digest-based fingerprints are derived for each BAC clone. Pairwise comparisons of the fingerprints are made and the resulting information is analysed to assume clone overlaps, which in turn are used to assemble BAC contigs. The BAC-fingerprinting process has been made relatively high-throughput in nature, allowing high-resolution BAC contig maps to be constructed for many plant species. Once provided with an assembled BAC contig map, minimally overlapping clones are then chosen for shotgun sequencing. For each selected BAC, the cloned DNA is purified and subjected to random fragmentation. The DNA fragments in a defined size range (usually 2-5 kb) are recovered and subcloned into a vector. Sequence reads are generated from both insert ends of a very large number of subclones, so as to produce highly redundant sequence coverage across the genome. The sequence reads are then computationally assembled on the basis of detected sequence overlaps. The resulting assembly usually produces a series of sequence contigs, each of which consists of a collection of overlapping reads and a deduced consensus sequence. After final assembly, the finished sequence is typically analysed for several important features, such as the presence and correct order of known sequence-based. Such authentication checks are crucial for ensuring that the final sequence produced in a shotgun-sequencing project is highly accurate.

To date, a number of plants are being sequenced using the clone-by-clone strategy, such as maize (Rabinowicz & Bennetzen, 2006), rice (Burr, 2002; Sasaki *et al.*, 2005), *Medicago truncatula* (Young *et al.*, 2005) and *Lotus japonicus* (Young *et al.*, 2005; Sato & Tabata, 2006).

1.8.2 Whole-genome shotgun sequencing

An alternative strategy for genome sequencing, called whole-genome shotgun sequencing, involves the assembly of sequence reads generated in a random, genome wide fashion (Figure 1.6). This approach has been attractive because the actual sequencing can be accomplished in a short period of time. It can easily be scaled because all clones can be sequenced in parallel.

In essence, the entire genome of an organism is fragmented into pieces of defined sizes, which in turn are subcloned into plasmid vectors. Sequence reads are generated from both insert ends of a very large number of subclones, so as to produce highly

redundant sequence coverage across the genome. Computational methods are then used to assemble the sequence reads and to deduce a corresponding consensus sequence. While this approach works well for genomes with a low content of repeat sequences, it is more challenging for genomes with a high content of repeat sequences. Therefore, a key aspect of this strategy, which is especially important for dealing with the problems presented by repetitive sequences, is the generation of sequence reads from both ends of most subclones (Edwards *et al.*, 1990). The major challenge, however, is the assembly of these sequences into contiguous sequence information and to anchor them to the genetic map. The expected physical distances separating these juxtaposed read pairs are an important factor in the process of deriving an accurate sequence assembly.

The whole-genome shotgun approach was first proposed by Craig Venter and colleagues as a means of speeding up the acquisition of contiguous sequence data for large genomes such as the human genome and those of other eukaryotes (Venter *et al.*, 1998; Marshall, 1999). Experiences with whole-genome shotgun sequencing have also revealed important advantages and disadvantages of the approach. Advantages include the ability to instigate the sequencing of a genome without an existing clone-based physical map and to produce large amounts of sequence data from an entire genome relatively quickly for use in identifying conserved sequences and polymorphisms. The main problems with the strategy, especially when used in the absence of supplementary clone-derived sequence data, consist of the sequence gaps and misassemblies that are caused by repetitive sequences (particularly those associated with low-copy duplicated segments) and the ambiguity about how best to proceed from data generated exclusively by whole-genome shotgun sequencing to a highly accurate, finished sequence of an entire genome.

In summary both sequencing strategies allow the generation of a significant amount of information. A whole-genome shotgun-sequencing strategy is particularly appropriate for studying a genome in a global and survey-orientated fashion, especially when a finished genome sequence is not an immediate goal. By contrast, clone-by-clone strategy is typically used when a more detailed, highly accurate (most likely finished) genome sequence is desired In addition, the clone-by-clone strategy can be scaled down. Such targeted sequencing, which usually involves the shotgun

sequencing of mapped BACs, provides the capacity to directly compare orthologous sequences generated from numerous different organisms. So, in addition to comprehensive genome sequencing for a select set of organisms, the generation of sequence from the same targeted genomic regions of an even larger set of organisms that will be used for detailed comparative sequence analysis.

1.9 Comparative structural genomics

Sequencing the genomes of the human, the mouse and a wide variety of other organisms is driving the development of an exciting new field of biological research called comparative genomics. By comparing the human genome with the genomes of different organisms, researchers can better understand the structure and function of human genes and thereby develop new strategies in the battle against human disease. In addition, comparative genomics provides a powerful new tool for studying evolutionary changes among organisms, helping to identify the genes that are conserved among species along with the genes that give each organism its own unique characteristics.

Over the past 20 years, plant comparative genetics has shown that the organisation of genes within genomes has remained more conserved over the evolutionary periods than previously thought (Gale & Devos, 1998). One of the virtues of comparative genomics lies in the transfer of structural and functional information from one genome to another (Vandepoele *et al.*, 2002). Homologies between the genomes of two species can be investigated either at the macro- or microsyntenic level depending on the available data. Macrosyntenic studies focus on the genomes as a whole analysing large regions (e.g., linkage groups) by comparing the order of the genes based on the constructed genetic maps. Microsyntenic comparisons use shorter but continuous stretches of completely sequenced genomic regions in which the order and the orientation of coding sequences as well as the non-coding DNA sections can be investigated.

1.9.1 Macrosynteny studies

Macrosynteny generally refers to conserved gene order between species revealed by comparative genetic mapping of common DNA markers or in silico mapping of homologous sequences. Comparative genomic mapping has demonstrated that plants have retained different levels of conservation in their genomes during evolution depending on their phylogenetic separation (Paterson *et al.*, 2000; Choi *et al.*, 2004a; Choi *et al.*, 2004c; Zhu *et al.*, 2005). These conserved regions, so called syntenic or orthologous regions, have collinear gene contents when compared either genetically or physically. The main plant taxa that have been extensively studied at macrosyntenic level include the Brassicaceae (crucifers), the Solonaceae, the Poaceae (grasses) and the Fabaceae (legumes).

The first comparative genetic mapping experiments in plants were performed on members of the Solanaceae family. The Solanaceae include several economically important plant species such as potato, tomato, pepper, eggplant and tobacco. Comparative genome analysis is a well-developed area in the Solanaceae. Early work showed that the tomato and potato genomes differ only five paracentric inversions (*i.e.*, inversions did not involve the centromere) (Bonierbale *et al.*, 1988; Tanksley *et al.*, 1992) but that the tomato and pepper genomes differ by numerous rearrangements (Tanksley *et al.*, 1988). Comparison of linkage maps of eggplant and tomato revealed the number and types of rearrangements that occurred during the evolution of these two species from a common ancestor (Bonierbale *et al.*, 1988).

A remarkable degree of genome conservation has been established in comparative genetic mapping experiments for the Poaceae family, although genome sizes vary as much as 40-fold between some of the species, and despite the fact that they diverged as long as 60 million years ago. This cereal family includes rice, wheat, maize, barley, sorghum, sugar cane, oat, rye, millet and others. The first publications that demonstrate the extent of synteny among the cereals showing the relationship between rice and maize (Ahn & Tanksley, 1993), that the genome of wheat could be aligned with rice (Kurata *et al.*, 1994) and Moore et al showed that all the maps could be combined in a single analysis (Moore *et al.*, 1995). A couple years after, the extent of the synteny within the Poaceae was represented in an integrated grass genome map, which includes species belonging to 6 different tribes and 3 different

subfamilies (Devos & Gale, 1997). The alignment in Figure 1.7 is based on rice, with the other genomes arranged relative to rice in the most parsimonious manner. Rice is used as the base simply because it is the smallest cereal genome analysed in detail, and for which the densest maps and most genomic tools are available.

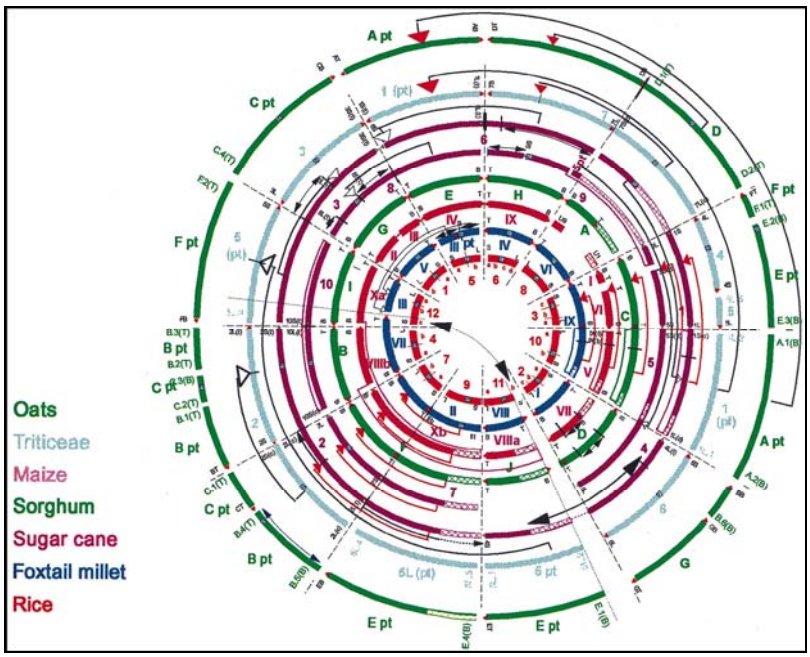


Figure 1.7. Aligned maps of rice, foxtail millet, sugar cane, sorghum, maize, the Triticeae crops and oats (Devos & Gale, 1997).

In the Brassicaceae family, the *Brassica* species have been the subject of many comparative genetic mapping experiments, not least because of their close relationship to *Arabidopsis thaliana*. At least 15 molecular genetic maps have been constructed for *Brassica* that include all of the major cultivated species (Kowalski *et al.*, 1994; Lagercrantz & Lydiate, 1996; Lagercrantz *et al.*, 1996; Lan *et al.*, 2000; Town *et al.*, 2006). These maps include a large number of publicly available probes, with many shared by multiple maps, and an almost complete conservation of gene repertoire was found between species. Within a species, the maps show almost complete colinearity, except for small inversions (Lan *et al.*, 2000). Comparison of *B. rapa* with *B. oleraceae* and *B. napus* supports the close evolutionary relationship between the two diploids but indicates that deletions and insertions may have occurred after divergence of the two species (Lagercrantz & Lydiate, 1996).

Early comparative studies of legume genomes were focused on closely related species of the same genus or tribe, based primarily on comparative mapping of common RFLP markers. The first study reported conserved gene order between pea and lentil, accounting for approximately 40% of the lentil genome (Weeden *et al.*, 1992). Later, another report demonstrated that mungbean and cowpea (*Vigna unguiculata*) also exhibited a high degree of linkage conservation, whereas chromosomal rearrangements have occurred since the divergence of the two species (Menancio-Hautea *et al.*, 1993). Comparative mapping among mungbean, common bean, and soybean in the Phaseoleae tribe indicated that mungbean and common bean linkage groups were highly conserved, but synteny with soybean was limited only to the short linkage blocks (Boutin *et al.*, 1995). A more recent study, however, using *Arabidopsis thaliana* as a bridging species revealed that homoeologous segments of soybean chromosomes showed a higher degree of synteny with chromosomes of common bean and mungbean than previously thought (Lee *et al.*, 2001). The most in-depth analysis of legume macrosynteny recently was reported using *Medicago truncatula* as a central point of comparison (Choi *et al.*, 2004a; Choi *et al.*, 2004c). A simplified consensus comparative map of eight legume species is shown in Figure 1.8. As expected, the degree of synteny is correlated with the phylogenetic distance of these legume species. *M. truncatula* and alfalfa share highly conserved nucleotide sequences and exhibit nearly perfect synteny between the two genomes (Choi *et al.*, 2004a). Although the pea genome is approximately 10 times larger than that of *M. truncatula* and has one less chromosome, the colinearity of genes is also extremely conserved between the two genomes, with major evident differences being inferred interchromosomal rearrangements.

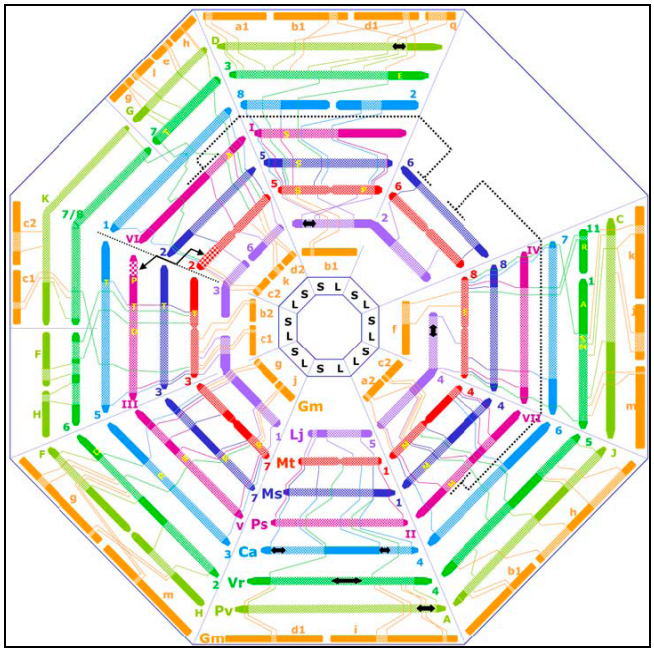


Figure 1.8. A simplified consensus map for eight legume species. Mt, *M. truncatula*; Ms, alfalfa; Lj, L. japonicus; Ps, pea; Ca, chickpea; Vr, mungbean; Pv, common bean; Gm, soybean. S and L denote the short and long arms of each chromosome in *M. truncatula*. Syntenic blocks are drawn to scale based on genetic distance (Zhu *et al.*, 2005).

1.9.2 Microsynteny studies

Comparative genetic mapping has generally revealed collinear chromosomal segments in closely related plants, whereas comparative genome studies at the nucleotide sequence level have disclosed many small differences between genomes of closely related species. As well as having ramifications in the ease with which comparative genomics information can be used for gene discovery, studies at this level also reveal much about genome evolution in plant species.

Over the last decade, a number of reports have studied the level of microsynteny between the grasses. The first comparative study of local gene content and order in the grasses revealed that several genes are conserved in order and orientation in *sh2/a1*-homolgous regions of the maize, rice and sorghum genomes (Chen *et al.*, 1997; Chen *et al.*, 1998). They have first discovered a duplication of a1 homologues in sorghum, separated by 10 Kb. In maize, *sh2* and *a1* were approximately 140 Kb apart and transcribed in the same direction, with *sh2* upstream of *a1*. In rice and sorghum, this arrangement is fully conserved; however the *sh2* and *a1* homologues were separated by about 19 Kb in both species (Figure 1.9) (Chen *et al.*, 1997). A more detailed comparison of this region between rice and sorghum indicated

conservation of gene presence and order of four loci (*sh2*, *a1*, X1 and X2) (Chen *et al.*, 1998). Interestingly the first exon of X2 in the rice locus is absent in the maize and sorghum genes. These studies reflect the fact that, while gene content and order are frequently conserved between species, genome size (and thus intergenic distances) is subject to massive variation. The maize genome is organised into 10 chromosomes and is about 2500 Mb in total. Sorghum, which is estimated to have diverged from a common ancestor with maize about 15-20 million years ago (MYA), has the same chromosome number, but its genome is about one third of the size. Rice diverged from a common ancestor with maize and sorghum about 50-60 MYA and has 12 chromosomes, comprising a much smaller genome of about 430 Mb. The size differences of the genomes are presumed to result from the ancestral allotetraploidization of the maize genome (Gaut & Doebley, 1997) and differences in the expansion and dispersion of repetitive DNA (White & Doebley, 1998).

More recently, Li and Gill demonstrated that the only X2 homolog in wheat is not linked to a X1 homolog (Li & Gill, 2002). Figure 1.9 shows that the wheat X1 and *sh2* homologs are linked, as are those of others grasses, but at a much greater distance that in other grasses. Similarly, the X2 and *a1* homologs are linked in wheat, but are located on a different chromosome from X1 and *sh2*. A translocation point appear to have occurred between X1 and X2 in an early Triticeae because the same rearrangement was observed in barley (Li & Gill, 2002). Thus, microsynteny must be viewed in the larger context of macrosyntenic relationships between species.

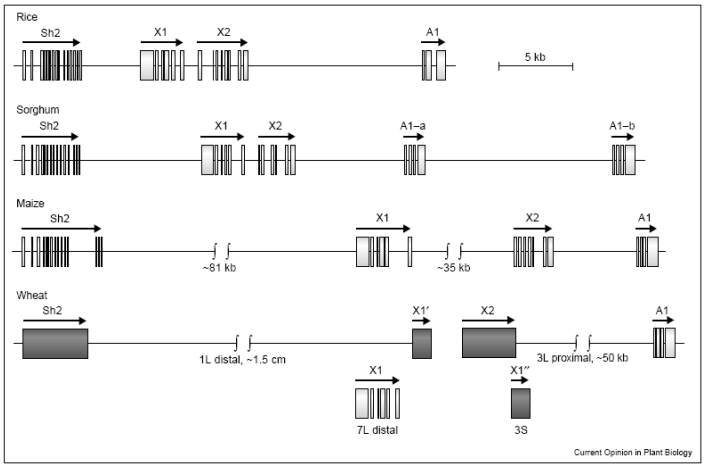


Figure 1.9. Restriction maps of regions *sh2* and *a1* homologues from maize, rice, sorghum and wheat (Bennetzen & Ma, 2003).

A second set of cereal chromosome segments that have been extensively studied by sequence analysis is the *alcohol dehydrogenase1* (*adh1*)-orthologous regions of maize, sorghum and rice (Tikhonov *et al.*, 1999; Tarchini *et al.*, 2000). Nine candidate genes including adh1 were found in a 225 Kb maize sequence, whereas in a 78 Kb space of sorghum the same nine genes were identified in a colinear order, plus five additional genes (Tikhonov *et al.*, 1999). Comparison of the rice *adh1/adh2* region with the sequence of the same region from maize suggested that *adh1* was transposed as a single gene to a new location (Tarchini *et al.*, 2000). However no change in the basic exon/intron organisation or alteration of *ahd1* expression resulted from this transposition (Tarchini *et al.*, 2000).

Comparative genomics (at the microsyntenic level) of *Arabidopsis thaliana* and *Brassica* species has been useful in understanding not only the evolution of genomes within the Cruciferae, but also of plant genomes in general. One of the earliest studies of this nature involved the comparison of *A. thaliana* and *B. oleracea* (Kowalski *et al.*, 1994). The authors discovered eleven regions of conserved organisation, spanning 24.6% of the *A. thaliana* genome and 29.9% of the *B. oleracea* genome. They detected at least 17 translocations and 9 inversions distinguishing the two genomes. This study also gave some insight into the fact that *Arabidopsis*, once thought of as the ultimate example of a diploid species, may also have undergone a round of whole genome duplication during its evolution. Some chromosomal segments in *B. oleracea* appeared to be in triplicate and the distribution of duplicated loci in *A. thaliana* suggests that ancient duplication may have occurred in Arabidopsis. This event was further emphasised in another study, analysing a 19 cM region in the *B. nigra* genome (Lagercrantz *et al.*, 1996). All of the *A. thaliana* sequences tested were present in three duplicate copies and with only one rearrangement disrupting perfect *Brassica/Arabidopsis* colinearity. Analysis of the whole genome sequence of Arabidopsis has subsequently revealed that the genome of this species has undergone not just one, but possibly two or three rounds of whole genome duplication, leading to a realisation that, even comparisons to apparently simple model genomes must be treated with care.

Conserved microsynteny is by no means ubiquitous, even in gene-containing regions of the genome. An in-depth study was reported recently based on the comparative

analysis of a transposon-rich *B. oleracea* BAC clone and its corresponding sequence in *A. thaliana* (Gao *et al.*, 2005). The *B. oleracea* BAC clone contained 8 genes and 5 transposable elements. Comparison with the *A. thaliana* showed that the first two genes were found at the end of chromosome V in *A. thaliana*. The third gene corresponded to an ortholog at the opposite end of the same chromosome, while the other 5 genes had a corresponding region on the same chromosome but further away (7.7 Mb).

Like other plants, legumes are likely to have a polyploid origin. Soybean, for example, has long been known to be an ancient polyploid with putative homoeologous chromosomal regions readily identified by genetic mapping (Shoemaker *et al.*, 1996). Evidence of ancient segmental duplications has also been found for the *M. truncatula* genome (Yan *et al.*, 2003; Zhu *et al.*, 2003), suggesting that its ancestor may have sustained largescale duplication event(s). However the timing of whole genome duplication events remains in dispute (Blanc & Wolfe, 2004).

With the concept of *M. truncatula* as a model legume with large-scale genome sequencing, a growing number of studies have begun to reveal extensive microsynteny between the members of the legume family. A study showed that conserved synteny is widespread between *M. truncatula* and soybean. Three homologous BAC contigs groups were analysed in detail and six of eight regions exhibited conserved synteny, including three that were extensively conserved (Yan *et al.*, 2004). Another report detailed the preliminary examination of the soybean *rhg1* region (Choi *et al.*, 2004c), where nearly 70% of genes were conserved and collinear between soybean and *M. truncatula*. Due to the genome sequence availability of *M. truncatula* and *L. japonicus*, a more in depth analysis of similarity between these two species has been carried out (Cannon *et al.*, 2006a). This study showed the existence of 10 large-scale synteny blocks, which account for 67% of the *M. truncatula* genome and 64% of *L. japonicus*. Within each synteny blocks, conservation of gene content and order was averaged to 54% (Cannon *et al.*, 2006a).

1.10 Context and objectives

In summary, comparative genomics studies confirm the idea that conserved genome synteny can facilitate transfer of knowledge among related species of plants. Comparisons of species that are phylogenetically more distant reveal that syntenic regions are often reduced to short linkage blocks and the degree of synteny declines with increasing phylogenetic distance, but even relatively closely related species can experience breakdown in conservation of gene order, which can make the above goal difficult.

The study described in this thesis seeks to provide a preliminary assessment of the levels of conserved synteny (particularly on a microsyntenic level) between white clover and the model species *Medicago truncatula*. The project involves the construction of a genetic map of white clover (*Trifolium repens* L.) using different types of molecular markers such as AFLPs, white clover SSR and some markers previously mapped in *Medicago truncatula.* The project also includes the construction and preliminary characterisation of a BAC library of one of the white clover mapping parents. A strategy to identify orthologous regions of clover and *Medicago truncatula*, based on BAC-end sequencing is proposed and validated. The genetic map and BAC library constitute a set of genetic and genomic resources which provide tools to assay the degree of synteny between white clover and *Medicago truncatula* with the ultimate aim of realising the potential of a comparative genomic approach for gene discovery and isolation in white clover.

2.0 Construction of a genetic linkage map of white clover and quantitative trait loci (QTL) analysis for morphological traits

2.1 Introduction

Molecular marker-based genetic linkage maps have been developed in a number of important plant species such as *Arabidopsis thaliana* (Chang *et al.*, 1988; Hauge *et al.*, 1993), *Lotus japonicus* (Hayashi *et al.*, 2001b; Sandal *et al.*, 2002), *Medicago truncatula* (Kulikova *et al.*, 2001), rice (*Oryza sativa* L.) (Chen *et al.*, 2002), and many more. Linkage maps are a fundamental tool for the identification of the genetic determinants underlying the expression of agronomically important traits. Genetic maps can be obtained through the use of DNA-based markers. The characteristics that determine the usefulness and value of a particular DNA marker class are its informativeness, reproducibility, cost, quality and quantity of the required DNA, transferability to a number of populations within and among species, and also the need for specific genomic probes (Lotti *et al.*, 2000).

In the legume family, the development of genetic linkage maps has been increasing in the major species such as soybean (Song *et al*., 2001), azuki bean (Han *et al*., 2005), cowpea (Menendez *et al*., 1997; Ouedraogo *et al*., 2002), lentil (Eujayl *et al*., 1998), chickpea (Cobos *et al*., 2005). In addition several whole-genome linkage maps in Trifolieae taxa have been developed, in diploid and tetraploid *Medicago sativa* (Brouwer and Osborn 1999, Kalo *et al*., 2000, Julier *et al*., 2003) and the model species *M. truncatula*. A restriction fragment length polymorphism linkage map of red clover (*T. pratense*) has been developed in a backcross population (Isobe *et al*., 2003), followed by the first molecular marker-based genetic map of white clover (Jones *et al.*, 2003). This map was a framework genetic map of white clover using genomic microsatellites (SSRs) in addition to amplified fragment length polymorphims (AFLPs) in an F_2 population derived from a pair cross between partial inbreds. In 2004, Barrett *et al.* constructed a more extensive genetic linkage map of white clover using the combination of in silico white clover EST-SSRs and genomic SSRs.

Genetic map-based analysis permits the dissection of complex phenotypes by resolving the location of interacting and pleiotropic genetic factors. Understanding the genetic basis of morphological change is of particular importance to the field of plant taxonomy because morphological characters provide the foundation for species description and play a central role in systems of classification. There have been

relatively few reports to date of QTL analysis for agronomic and morphological traits in white clover. Following the development of a comprehensive genetic linkage map of white clover, Barrett *et al*. used the resource to identify QTLs that regulate seed yield and three of its components, inflorescence density, yield per inflorescence and thousand-seed weight (Barrett *et al*., 2005a). In another study, they analysed the genetic location of a root-knot nematode resistance locus in Trifolium using bulk segregant analysis (Barrett *et al*., 2005b). Recently, trait dissection and QTL identification from multiple environments for white clover was carried out on the basis of the F_2 (I.4RxI.5J) population used for the construction of the genetic map from Jones *et al*. (2003) (Cogan *et al*., 2006).

This chapter details the construction of a genetic linkage map of an F_1 (R3R4 x S1S4) white clover population. This map is composed of AFLP markers and SSR markers previously mapped in white clover (Jones *et al.*, 2003; Barrett *et al.*, 2004) and also SSRs that were developed from the analysis of a bacterial artificial library of the one of the parent material (R3R4) (Febrer *et al.*, 2007). In the context of this thesis, the genetic map will not be used as a platform for comprehensive comparative genetic mapping between white clover and *Medicago*, but rather as a tool to enable the sequence-based information developed in the subsequent two chapters to be placed in a positional genetic context.

This chapter also details a phenotypic analysis carried out both in a glasshouse and a field experiment. Morphological traits such as leaf length and width, internode length and diameter, petiole length, plant height and spread, and flowering time were assessed and quantitative trait loci (QTL) for morphological traits were detected in the genetic linkage map of white clover.

2.2 Materials and Methods

2.2.1 Plant material

An F_1 tetraploid white clover (*Trifolium repens* L.) population comprising 94 progeny individuals, produced in IGER (Wales, UK) by crossing the white clover genotypes R3R4 and S1S4, was used as mapping population.

2.2.2 Genomic DNA isolation

2.2.2.1 DNA isolation using CTAB method

Approximately 2 g of fresh leaf material was ground using liquid nitrogen and 10 ml extraction buffer [0.1 M Tris-HCl (hydroxymethyl aminomethane hydrochloride, Sigma™) pH 7.5; 1.4 M NaCl (sodium chloride, Sigma™); 0.05 M EDTA (ethylenediamine tetracetic acid, Sigma™) pH 8.0; 2% CTAB (hexadecyltrimethylammonium bromide, Sigma™); 0.2 % β-mercaptoethanol] was added. After incubation at 65°C for 1 hour, 5 ml of chloroform/isoamyl-alcohol (24:1 v/v) were added followed by centrifugation at 10,000 rpm for 15 min. The aqueous layer was removed and genomic DNA precipitated by addition of an equal volume of chilled (4°C) isopropanol. The precipitated DNA was recovered either by centrifugation or by "spooling" on a glass hook. The DNA was washed in 70% ethanol and dried over night before being resuspended in 500 µl TE buffer [10 mM Tris (hydroxtmethyl methylamine, Sigma™), 1 mM EDTA, pH 8.0].

2.2.2.2 Estimation of DNA concentration

DNA concentration was estimated by electrophoresis on a 1% agarose gel run in 1X TBE (89 mM Tris pH 8.3, 89 mM boric acid, 2 mM EDTA) alongside with known quantities of λDNA. Gels were run at 100 Volts (V) for 1 hour and DNA was visualised by staining with an ethidium bromide solution (0.5 µg/ml) and viewed on an ultra violet (UV) transilluminator (Kodak Digital Science, Image station 440CF).

2.2.3 AFLP Procedure

2.2.3.1 Restriction-Ligation of the genomic DNA

0.25 µg of genomic DNA was fully digested for 1 hour at 37°C using 2.5 units *Pst*I (New England Biolabs), 2.5 units *Mse*I (New England Biolabs) and 1X buffer 2 (New England Biolabs), 0.1 mg/ml BSA (New England Biolabs). After digestion, 5

µl ligation mixture containing 2.5 pmol *Pst*I or *Eco*RI adapter (Table 2.1), 2.5 pmol *Mse*I adapter (Table 2.1), 2.5 µl 10X ligase buffer which contains 10 mM ATP and 0.5 unit T4 DNA ligase (5 units/µl, New England Biolabs) were added. The template was then incubated for a further 3 hours at 37°C in a Peltier thermal cycler 200 (MJ Research, USA).

Table 2.1. Sequences of the adapters used for the ligation of the restricted genomic DNA.

***Pst*I adapter**	
Forward	5' – CTCGTAGACTGCGTACATGCA – 3'
Reverse	5' – TGTACGCAGTCTAC – 3'
***Eco*RI adapter**	
Forward	5' – CTCGTAGACTGCGTACC – 3'
Reverse	5' – AATTGGTACGCAGTCTAC – 3'
***Mse*I adapter**	
Forward	5' – GACGATGAGTCCTGAG – 3'
Reverse	5' – TACTCAGGACTCAT – 3'

Both adapters were prepared by mixing the bottom and top strands (3000 pmoles of each strand) and sterile distilled water to obtain stocks of 50 pmol/µl of *Mse*I adapter and 5 pmol/µl of *Pst*I or *Eco*RI adapter. The adapters were denatured at 65°C and cooled down slowly to room temperature to optimise annealing.

2.2.3.2 Non-selective amplification (pre-amplification)

The purpose of this intermediate step is to generate a large amount of secondary template DNA to perform the AFLP reaction with radioactively labelled selective primers. The primers used in the pre-amplification are non-selective primers (Table 2.2), strictly complementary to the adapters.

Table 2.2. Sequences of the non-selective primers used for the pre-amplification step of the AFLP procedure.

	CORE	ENZYME
MseI primer (M_{00} primer)	5'-GATGAGTCCTGAG	TAA-3'
EcoRI primer (E_{00} primer)	5'-GACTGCGTACC	AATTC- 3'
PstI primer (P_{00} primer)	5'-GACTGCGTACATGCA	G-3'

Pre-amplification PCR reactions were performed in a Peltier thermal cycler PTC-200 (MJ Research, USA) as described in Table 2.3 and 2.4. The amplification products were checked by electrophoresis on a 1% agarose gel in 1X TBE stained with ethidium bromide. The expected result was a smear ranging from 100-500 bp. This

secondary template was diluted depending on the intensity of the smear, (generally 25-fold) in $T_{0.1}E$ (10 mM Tris pH 8.0, 0.1 mM EDTA) and stored at -20°C.

Table 2.3. Pre-amplification PCR components.

Component	Volume/single reaction	Final concentration
Ligated DNA	5	2.5 ng/20 µl
10X Buffer	2	1X
dNTPs (2 mM)	2	200 µM
P_{00} primer (30ng/µl)	1	30 ng
M_{00} primer (30ng/µl)	1	30 ng
Taq polymerase (5 u/µl)	0.08	0.02 u/20 µl
dH_2O	8.92	To a final volume of 20 µl

Table 2.4. Amplification condition for Pre-amplification PCR.

Amplification:	94°C x 0.30min	
	56°C x 0.30min	24 cycles
	72°C x 1 min	
Hold temperature	4°C	

2.2.3.3 Selective amplification

For the selective amplification of a sufficient number of polymorphic fragments, a total of 30 primer combinations were used (Table 2.5). The P primer had a 2-base extension (P + 2) and the M primer and the E primer had a 3-base extension (M + 3, E + 3). They were chosen because in pilot experiments they gave the best combination of clarity and polymorphism.

Table 2.5. AFLP primer combinations for white clover population.

	Pac	Pcg	Pta	Pat	Pgc
Maac	X	X	X	X	X
Maag	X	X	X	X	X
Maat	X	X	X	X	X
Maga	X	X	X	X	X
Matc	X	X	X	X	X
	Eaac	Eaca	Eaag	Eacg	Etcg
Mcca	X	X	X	X	X

Note: The letters after each primer name are the nucleotide extensions added to each primer sequence.

For each reaction, 5 ng of the P + 2 or E + 3 primer was end-labelled with 1.0 µCi $\gamma[^{33}P]$-ATP, in 1X T4 buffer using 2 units/µl T4 polynucleotide kinase (New England Biolabs). The mixture was incubated at 37°C for one 1 hour and then the kinase was inactivated by incubation at 65°C for 10 minutes.

For the AFLP reaction, 5 µl of secondary template was mixed with 5 ng of γ[^{33}P]-ATP end-labelled P or E primer and 30 ng of unlabelled M primer, the other components were the same as Table 2.3. The PCR reactions were performed using a touch-down profile (Table 2.6).

Table 2.6. Touch-down PCR profile for selective amplification.

Amplification 1:	94°C x 0.30min	
12 cycles	65°C - 56°C x 0.30min	Decrease 0.7°C each cycle
	72°C x 1 min	
Amplification 2:	94°C x 0.30min	
24 cycles	56°C x 0.30min	
	72°C x 1 min	
Hold temperature	4°C	

2.2.3.4 Separation of labelled fragments and autoradiography

An equal volume of loading dye [98% de-ionised formamide (Sigma), 10 mM EDTA pH 8.0, bromophenol blue (1 mg/ml), xylene cyanol (1 mg/ml)] was added to the reaction products from 2.3.3. 3 µl of amplification products were loaded on a 5% acrylamide gel (5% polyacrylamide, 1X TBE, 75 M Urea, 10% APS, 100 µl Temed) using a BioRad electrophoresis system (Richmond, VA, USA). 1X TBE was the loading buffer used. The run was performed at a constant 110 W for 2.5 hours. The gels were dried on Whattmann 3MM paper, and exposed to storage phosphor screens for 1 to 3 days at room temperature.

2.2.4 Microsatellites (SSRs)

2.2.4.1 Source of SSRs used in this study

In order to provide chromosomal identification and orientation, to identify homologous and homoeologous groups and to anchor the map in this study to previously published maps, SSR markers from various sources were used.

- **Published SSR primer sequences**

Barrett *et al*. (2004) developed SSR genetic markers for the white clover genome by mining an expressed sequence tag (EST) database and by isolation from enriched genomic libraries. From the 30 genomic SSR and 32 EST-SSR published, 39 were chosen for this study to obtain a density of at least two SSR are on each linkage group (Appendix A).

In 2003, Jones *et al*. generated a set of c. 400 unique SSR clones from white clover, 100 of which were characterised for amplication and detection for polymorphism. These white clover SSRs (TRSSR) were used to create a molecular marker-based genetic map of white clover (Jones *et al.*, 2003). A small subset of those (Table 2.7) was used in order to anchor the SSRs markers from the published map on to the map in this study.

Table 2.7. List of *Trifolium repens* microsatellites (TRSSR) from Jones *et al*. (2003) used in the F_1(R3R4 x S1S4) mapping population.

TRSSR	SSR motif	Forward primer	Reverse primer	Map location[a]
A01C10	$(TC)_{17}(AC)_{16}$	GTACCTGGAAATGTTGATT	GAGCAGCCATGACCTCTG	6
A06B07	$(GT)_8$	TGTCAGATGTCATGCATATTTCAG	TTGAAGTGATTAACGAAGAAGGAC	9
A04F01	$(GT)_9\ldots(GT)_6$	TCCTTCGCCAGTCGTTTCAA	CGATCGTATCCTATGCTGTT	9
B01E07	$(GT)_{10}$	TTTTGTCTAATTGCAGAACCATGG	TTTAAGTAACAGGTTGATGCGTAC	13
B02E01	$(GT)_8(GC)_5$	ACGGGAGATAATTCATTTCTGAAG	GGTCGAGAAATACAACATGCATAC	16

[a] Map location according to Jones *et al*. 2003

- **Genbank derived SSRs**

At the time of writing, 450 genomic white clover sequences containing microsatellite motifs were publicly available at Genbank (www.ncbi.nlm.nih.gov/entrez). These sequences were identified via the Tandem Repeats Finder programme. Primer pairs were designed using the Primer3 software and tested subsequently for their quality of amplification (Table 2.8) (Barth *et al.*, 2004).

Table 2.8. List of Genbank white clover microsatellites used in the F_1(R3R4 x S1S4) mapping population (Barth *et al.*, 2004).

Marker alleles	Repeat motif	Size range	No loci
TRagrA1.28	$(CAT)_{10}$	111-242	40
Tragr1023	$(CA)_{17}$	150-265	28
Tragr1084	$(GT)_{18}$	121-236	27
TRagr179	$(ATCT)_7$	130-244	32

- **White clover BAC derived SSRs**

Following the construction of a bacterial artificial chromosome library of white clover (Febrer *et al.*, 2007), 34 microsatellites were designed from the BAC-end sequencing analysis (Chapter 3, Section 3.2.5) and 10 microsatellites were developed from six-fold coverage BAC clone sequencing analysis (Chapter 4, Section 4.2.3). All SSRs were analysed as previously mentioned and mapped onto the white clover genetic map.

2.2.4.2 Methods for microsatellite analysis

- **Initial testing**

All SSRs were first tested for amplification in the parental lines (S1S4, R3R4) and the PCR reaction was as described by Barrett *et al*. (2004) (Table 2.9 and 2.10).

Table 2.9. Standard PCR components for the test amplification in the parental lines for all SSRs.

Component	Volume/single reaction	Final concentration
DNA sample (25ng/µl)	2	2.5 ng/20 µl
10X Buffer	2	1X
dNTPs (2 mM)	2	200 µM
Upstream primer (10 µM)	0.4	0.2 µM
Downstream primer (10 µM)	0.4	0.2 µM
Taq polymerase (5 u/µl)	0.25	0.05 u/20 µl
dH_2O	12.95	To a final volume of 20 µl

Table 2.10. Amplification conditions used for the PCR reaction.

Initial Denaturation	94°C x 4 min	1 cycle
Amplification	94°C x 0.30 min 55°C or 50°C x 0.30 min 72°C x 1 min	30 cycles
Final extension	72°C x 7 min	1 cycle
Hold temperature	4°C	

Following the initial testing of the SSRs, the markers from the various sources were analysed either using radioactive labelling (SSRs from Barrett *et al*. 2004) or using fluorescent labelling (SSRs from Jones *et al*. 2003, Genbank derived SRRs, and white clover BAC derived SSRs).

- **Radioactive labelling**

Once tested for amplification, the SSRs from Barrett *et al*. (2004) were tested for polymorphism on the mapping population. This experiment was carried out using radioactive labelling with [^{33}P]-ATP (GE Healthcare). For each reaction the forward primer was end-labelled using 0.8 mCi/ml [^{33}P]-ATP (GE Healthcare), 1 X Optikinase Reaction Buffer (GE Healthcare), 0.2 units/µl Optikinase (GE Healthcare), 4 µM forward primer to a final volume of 0.5 µl. The labelling reaction was incubated at 37°C for 1 hour. PCR reactions were performed in a Peltier thermal cycler 200 (MJ Research, USA) as described in Table 2.11 and 2.12. The separation of labelled fragments and autoradiography were performed as described in Section 2.2.3.4.

Table 2.11. Components for radioactive labelled PCR.

Component	Volume/single reaction	Final concentration
DNA sample (25 ng/µl)	1	2.5 ng/10 µl
10X Buffer	1	1X
dNTPs (2 mM)	1	200 µM
[^{33}P]-ATP labelled primer	0.25	0.25 µM
Non labelled primer (10 µM)	0.1	0.1 µM
Taq polymerase (5 u/µl)	0.05	0.025 u/10 µl
dH_2O	6.6	To a final volume of 10 µl

Table 2.12. Amplification condition for radioactive labelled PCR.

Initial Denaturation	94°C x 3 min	1 cycle
Amplification:	94°C x 0.45min	
	50°C x 0.45min	34 cycles
	72°C x 0.45min	
Final extension	72°C x 2min	1 cycle
Hold temperature	4°C	

- **Fluorescent labelling**

The microsatellites analysed fluorescently were tested on the white clover mapping population using the ABI 3100 sequencer (Applied Biosystems). Each marker was fluorescently labelled with different colours using 4 fluorophors (FAM= blue, NED= yellow, VIC= green, PET= red). In that case, the PCR products can be pooled. The PCR reaction for each primer was set as described in Table 2.13 and 2.14. The plates were then incubated at 60°C for 30 minutes to avoid the formation of polyA peaks.

Table 2.13. Components for fluorescent labelled PCR.

Component	Volume/single reaction	Final concentration
DNA sample (25 ng/µl)	1	2.5 ng/10 µl
10X Buffer	1	1X
dNTPs (2 mM)	1	200 µM
Fluorescent labelled primer (10 µM)	0.1	0.1 µM
Non labelled primer (10 µM)	0.1	0.1 µM
Taq polymerase (5 u/µl)	0.1	1 u/10 µl
dH_2O	6.6	To a final volume of 10 µl

Table 2.14. Amplification condition for fluorescent labelled PCR.

Initial Denaturation	94°C x 5 min	1 cycle
Amplification	94°C x 1min 55°C x 1 min 72°C x 1 min	36 cycles
Final extension	72°C x 10 min	1 cycle
Hold temperature	4°C	

The PCR products were then pooled as follows: 1 µl of Fam, 1 µl of Vic, 2 µl of Ned and 2 µl of Pet. 0.5 µl of the pooled PCR products was added to 9.5 µl of formamide/sizer [25 µl size standard (GeneScan™ - 500 LIZ® Size Standard) + 950 µl formamide (Hi-Di™ Formamide1, ABI)] on an ABI 96-well plate. The plate was incubated at 95°C for 5 min and placed immediately on ice. The plate was then run on the ABI 3100 Genetic analyser and the data were analysed using ABI Prism® GeneMapper™ Software Version 3.0.

2.2.5 Linkage analysis and mapping

Markers with alleles segregating in one or both parents were analysed in the entire population of 94 F_1 progeny individuals. Both of the parents were used to define their respective coupling linkage groups (LGs) using JoinMap 3.0 (Van Ooijen & Voorrips, 2001). Construction of the linkage map was accomplished by treating the segregating data as a cross pollinator (CP). The two parental maps were first constructed with AFLP markers segregating uniparentally (with a 1:1 ratio-like pattern) and those that segregated biparentally (with 3:1 ratio-like pattern). The SSR markers were then included to identify the homologous and homoeologous groups in the two parents.

Dealing with co-dominant SSR markers in a supposed allotetraploid presented some difficulties due to the presence of two homoeologous genomes. A single SSR marker may actually amplify 2 essentially independent loci (both homoeologues), or alternatively only one of the two homoeologues present. Several factors can make it difficult to derive the actual marker genotype present at a locus from the banding pattern or chromatogram (marker phenotype).

These confounding factors include:

1. Independently segregating homoeologues may have one or more alleles of the same size.
2. Null alleles may be present at one or both of the homoelogous loci.
3. An extension of the above idea is that it is not inconceivable that the SSR marker is amplifying both homoeologues in one parent, but only one homoeologue in the other.

While it was possible to score SSRs in a co-dominant manner for some of the markers, based on observing the segregation patterns in the progeny, this was not possible for many of the loci. For consistency, each SSR allele was scored independently, in a dominant fashion, and linkage group identity, orientation and homoeology was established retrospectively by manually inspecting the assembled linkage groups. It is possible that some of the apparent SSR alleles scored in this manner were erroneous. A particularly likely error would occur where two different homoeologues in the different parents possessed a marker allele of the same size – this would be scored as a biparental marker with an expected segregation ratio of 3:1. Again, a retrospective approach was undertaken to eliminate such "synthetic false" alleles by examining the assembled linkage groups and removing any markers that did not fit well into the map.

2.2.6 Morphological measurements

In addition to the genetic map of white clover, a phenotypic analysis was carried out in order to further characterise the mapping population. This analysis was assessed in two different environments: a glasshouse experiment and a field experiment. Those two experiments were then compared to see the change in environment has an effect on the population. Two sets of cuttings from the mapping population were planted in a glasshouse in a randomised complete block design with four replicates each, one for the glasshouse experiment and the other for the field experiment (May 13th 2005).

2.2.6.1 Glasshouse measurements

All glasshouse measurements were assessed 56 days after the cuttings were planted (8-13 July 2005). The following measurements (Figure 2.1) were taken from the longest stolon for each plant:

1. Stolon length (SL). This measurement was taken from the tip of the primary stolon to as near to the base of the plant as possible.
2. Internode length (IL). This measurement was taken between the 3rd and 4th fully expanded leaf.
3. Internode diameter (ID). This measurement was taken behind the third fully expanded leaf from the apical tip.

4. Petiole length (PL). This was measured on the petiole associated with the third fully expanded leaf of the stolon.

5. Leaf size. The length (LL) and width (LW) of the middle leaflet on the measured petiole.

All measurements were taken in centimetres except for the internode length, which was measured in millimetres.

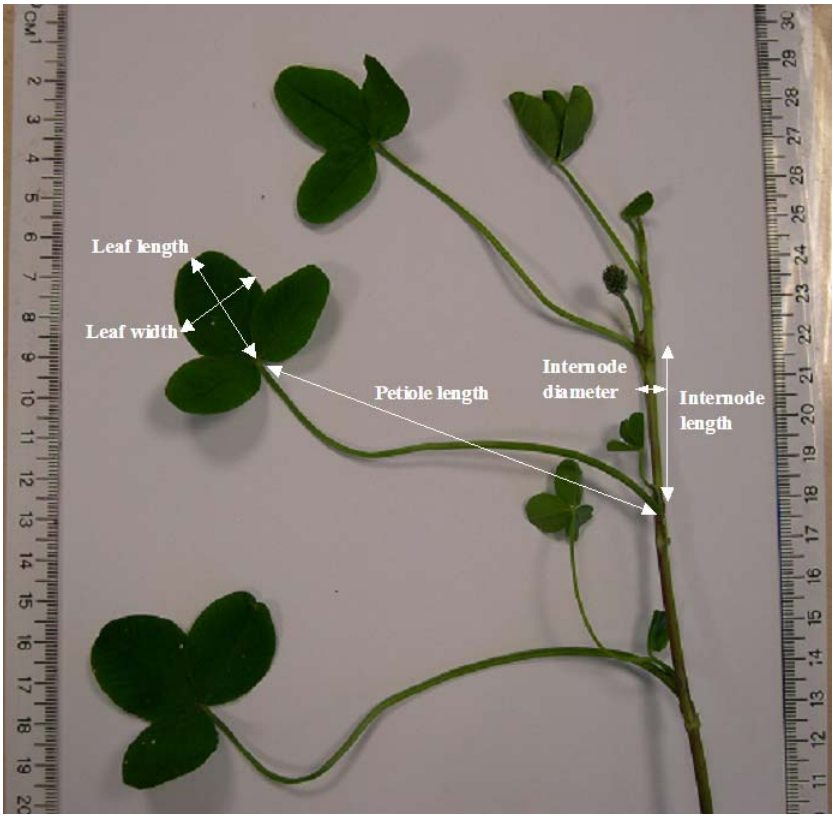


Figure 2.1. This picture represents a description of the measurements of the traits on the white clover plant.

2.2.6.2 Field measurements

The mapping family was also characterised in the field. The second set of cuttings that were grown in the glasshouse in May 2005, were planted in the field on August 24th 2005. The randomised complete block design was the same as the glasshouse one, with rows consisting of ten plants, plants and rows both at a spacing of 0.9 m. All plants in the field were measured in June 2006. The field measurements were carried out in a same as the glasshouse measurements with the exception of the stolon length (Figure 2.1). Additional measurements were also taken. These were as follows:

1. Plant spread. This was measured diagonally at the tips of the longest stolons on each side.
2. Plant height. This measurement was taken from the ground to the tip of the highest stolon.
3. Flower height. This measurement was taken from the ground on the highest flower of the plant.
4. Flowering date. This was measured every Tuesday and Friday starting from the April 1st 2006 until all plants flowered.

2.2.6.3 Data analysis

Means, standard deviations, coefficient of variation, analysis of variance (ANOVA) and Pearson correlation coefficients were calculated on the data collected using Minitab 14 statistical package.

2.2.6.4 QTL analysis

Single-marker regression (SMR) was initially employed to identify significant variation associated with the genetic markers used in the genetic map. Simple interval mapping (SIM) method was used to identify and confirm the presence of QTLs. All analyses were performed using MapQTL® 4.0 (Van Ooijen *et al.*, 2002). The maximum log-of-odds (LOD) score of association between the genotype and trait was calculated for SIM and QTL location predictions were accepted for values greater than a LOD threshold value of 2.5.

2.3 Results

2.3.1 AFLP markers

Twenty five *Pst*I/*Mse*I and five *Eco*RI/*Mse*I primer combinations were used to detect polymorphisms between R3R4 and S1S4 parental lines. The selected primer combinations (PCs) were highly polymorphic in the white clover population. A total 449 polymorphic bands were generated by the 30 primer combinations (60 by *Eco*RI/*Mse*I and 389 by *Pst*I/*Mse*I PCs) (Figure 2.2). The number of polymorphic loci per a primer combination ranged from 7 to 37, with the mean value of 16.75 (detailed in Table 2.15). The number of informative loci was slightly different between the *Eco*RI and *Pst*I assays, which produced in average 12 and 15.6 polymorphic fragments per PC, respectively. In addition, the profiles generated by the *Pst*I/*Mse*I PCs were clearer and easier to score than the *Eco*RI/*Mse*I PCs.

Of the 449 AFLP loci, 157 were polymorphic in R3R4 only, 216 were polymorphic in S1S4 only, and the remaining 76 segregated in both parental genotypes. A total of 233 and 292 AFLP loci were used to construct maps of R3R4 and S1S4, respectively. The number of biparental segregating AFLP loci was very low; therefore, in this mapping population, the use of AFLP markers to correlate the two parental maps was very difficult and needed the addition of information from SSR markers.

Table 2.15. Summary of the AFLP markers scored on the F_1(R3R4 x S1S4) mapping population.

	Pac	Pat	Pcg	Pga	Pta
Maac	26	12	13	11	14
Maag	31	10	15	19	15
Maat	37	9	9	16	7
Maga	21	12	17	19	10
Matc	19	19	10	19	23
	Eaac	Eaca	Eaag	Eacg	Etcg
Mcca	16	12	15	9	8

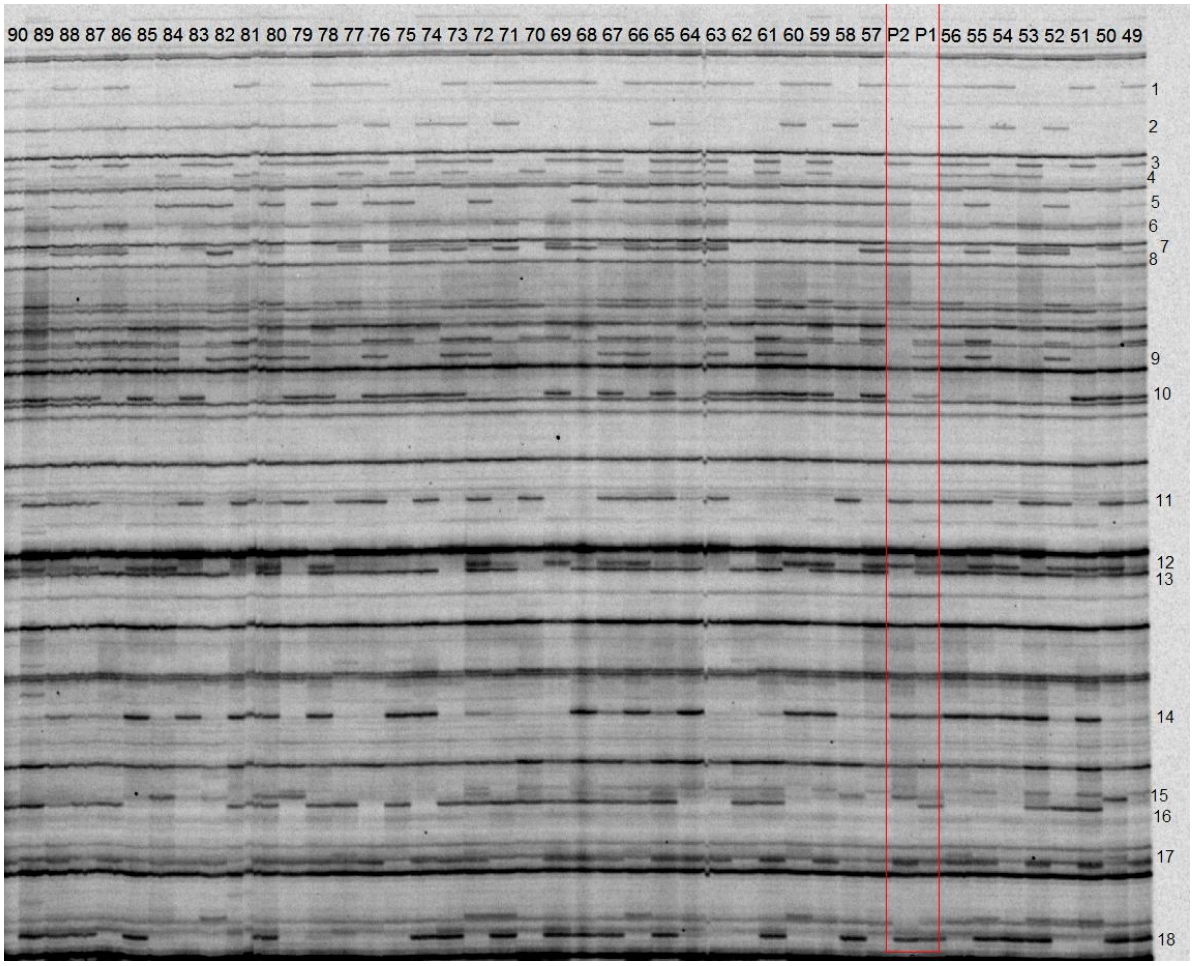


Figure 2.2. A 5% polyacrylamide gel showing the polymorphic bands of the PacMaag primer combination tested on the two parents and 42 progeny individuals (49 to 90).

2.3.2 SSR markers used

The five sets of SSRs were first tested for amplification in the parental lines (Figure 2.3).

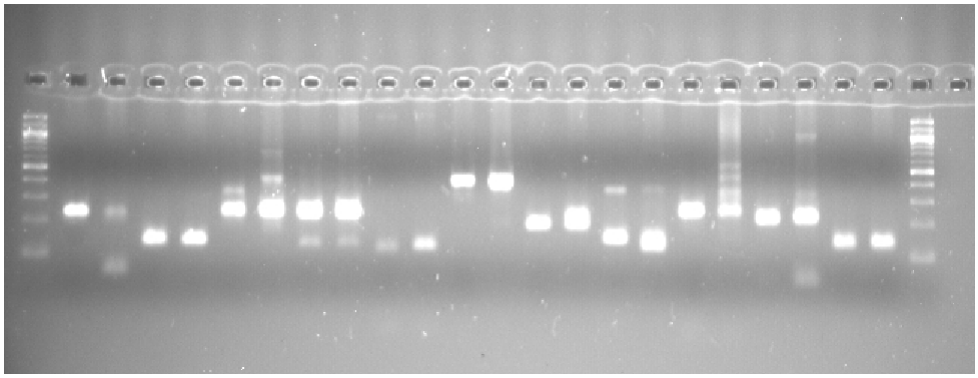


Figure 2.3. A 1% agarose gel illustrating the amplification of 11 white clover SSRs tested on the two parental lines (S1S4, R3R4).

The SSR primers that amplified in both parents were then applied on the whole mapping population and polymorphism was observed (Figure 2.4 and 2.5) and scored for the construction of the genetic map. A total of 59 SSRs were scored dominantly resulting 169 SSR alleles (Table 2.16). Of these 169 SSR alleles, 68 segregated from R3R4 only, 82 were polymorphic in S1S4 only, and the remaining 19 were segregating in both parental lines. A total of 87 and 101 SSR alleles were used to construct maps of R3R4 and S1S4, respectively.

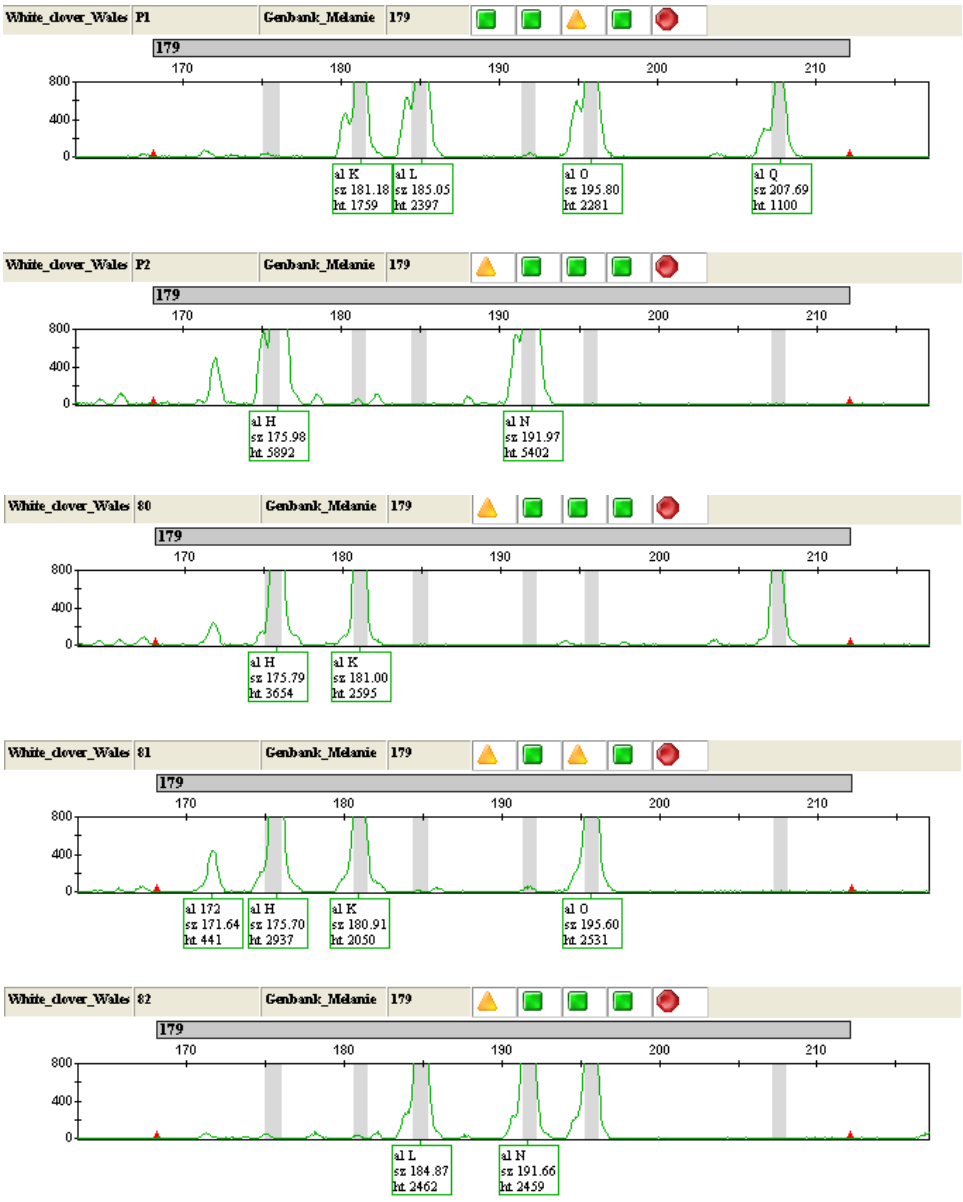


Figure 2.4. An ABI Chromatogram that shows the allelic pattern of a Genebank SSR (TrAgr179) on the parental lines and 3 progeny individuals. The shaded areas represent the possible alleles.

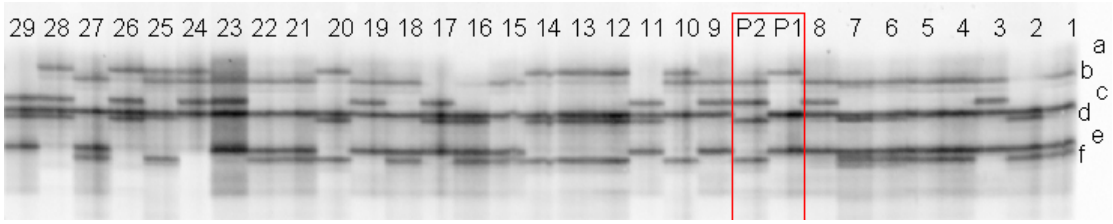


Figure 2.5. Polyacrylamide gel showing the polymorphism of a white clover SSR (ATS032) from Barrett *et al*. (2004) between the parents and 29 progeny.

Table 2.16. This table represents a summary of all SSRs used.

SSR type	Number of SSR	Amplification	Polymorphism	Scoring	Number of alleles
White clover SSR[a]	39	36	25	25	79
Genbank SSR	4	4	4	4	13
TRSSRs[b]	5	5	4	4	11
BES SSR[c]	34	32	24	18	44
BAC sequence SSR[d]	10	10	8	8	22

[a] SSRs developed by Barrett *et al*. 2004
[b] SSRs developed by Jones *et al*. 2003
[c] SSR developed from the BAC-end sequence analysis of the R3R4 BAC library
[d] SSR developed from the BAC clones that were fully sequences

2.3.3. Linkage map of white clover

The AFLP and SSR markers were used to construct the parental maps on the basis of segregation data obtained from 94 F_1 progeny (Table 2.17). From a total of 618 alleles segregating in the population, 85 % segregated from one parent and the remaining 15 % segregated from both parents. The nomenclature and orientation of each group was given according to previously published white clover linkage maps (Jones *et al.*, 2003; Barrett *et al.*, 2004).

The R3R4 parental map (Figure 2.6) was constructed with 167 AFLP loci and 55 SSR alleles, leaving 66 AFLP and 31 SSR alleles unlinked (Table 2.18). A total of 222 marker alleles were distributed in 15 linkage groups (LGs). The distribution of marker alleles by type is summarised in Table 2.19. Markers in the 15 LGs covered a length of 1138 cM. The average length of the linkage groups was 78.9 cM, with LGs ranging in length from 36 to 138 cM, and the average marker density in this parental map was 5.13 cM/marker.

The S1S4 parental map (Figure 2.7) was constructed with 220 AFLP and 81 SSR alleles, leaving 71 AFLP and 20 SSR alleles unlinked (Table 2.19). A total of 301 marker alleles were distributed in 22 linkage groups. The distribution of marker alleles by type is summarised in Table 2.20. Map length covered by all linkage

groups was 1494 cM, with the LGs ranging from 18 to 164 cM and an average of 67.9 cM. The average distance between markers was 4.96 cM/marker.

For the purposes of map alignment we have adopted the homoeologue group nomenclature of Barrett *et al*. (2004), who arbitrarily assigned a code (letters A-H) to each of the eight homoeologous pairs of LGs in clover. Barrett *et al*. subsequently differentiated each pair of homoeologues with a numerical suffix (A1, A2 – H1, H2). It is important to note that, unlike the situation in some other allotetraploids (such as wheat), there is currently no well defined differentiation between the two homoeologous genomes in white clover, and the numerical suffixes assigned by Barrett *et al*. (2004) are also arbitrary. Therefore the nomenclature for our map will be as follows: R-[1 to 15][(A to H)] or S-[1 to 20][(A to H)], where R and S represent the origin of the parental map (R3R4 or S1S4), followed by an arbitrary number for the linkage group, and finally the letter (A to H) which corresponds to the Barrett *et al*. (2004) homoeologue group classification.

For the parental map R3R4, the number of linkage groups observed was one short of the expected number of 16 LGs for white clover. Seven of the eight homoeologous pairs were successfully identified by SSR markers (missing pair H). Only one LG was identified for pairs B and G. Three linkage groups were assigned to pair A (54, 64, 68 cM). These cases in which a greater number of groups than expected were obtained probably represent instances where marker coverage was insufficient, resulting in an inability to detect linkage between groups of markers on the same LG. Two linkage groups (R-14(?) and R-15(?)) remain unidentified due to the fact that no previously mapped SSR was found on either LG. These may represent some of the missing LGs (B, G and H).

For the parental map S1S4, the number of LGs observed was 22, which is higher than the expected 16 LGs and the majority of the eight homoeologous pairs were successfully characterised except for one linkage group for pair F (Figure 2.7). Four linkage groups were assigned to the pair D, with two of a relatively small size (18 and 22 cM). Similarly, three linkage groups were assigned to pair H (31, 51 and 93 cM). Again, two linkage groups (S-19(?) and S-20(?)) were not successfully identified. The addition of more SSR markers on these linkage maps would help to

increase the density of the map and might eventually resolve the unequal number of linkage groups compared to the expected 16 linkage groups for white clover (2n = 4x = 32).

As previously mentioned the behaviour of a genetic marker from one segregating population to another may vary. In fact, it is clear from our results that while the SSR markers of Barrett *et al*. (2004) are consistently homoeologue specific, they are not necessary consistently homologue specific. For instance, some of the SSRs that were mapped on both homoeologues within a pair in the Barrett *et al*. map were only found on one homoeologue in the map presented here (e.g. ATS029 and ATS032 were present in A1 and A2 in the Barrett *et al*. map, but only on one A group here) and vice versa (e.g. ATS003, ATS113, PRS510).

In 2003, Jones *et al*. constructed another genetic linkage map of white clover using both AFLP and SSR markers. This map was assembled using a different nomenclature to those of Barrett *et al*. (2004) and was composed of unique SSR clones from white clover. Four SSR markers (TRSSR-A01C10, -A06B07, -B01E07 and -B02E01) were mapped in our mapping population. These markers were assigned to four separate linkage groups by Jones *et al*. (2003); these were linkage group 6, 9, 13 and 16 respectively, according to the nomenclature used in that study. In this study, two of the markers (TRSSR-A06B07 and -B01E07) were assigned to the two of the linkage groups B, suggesting that linkage group 9 and 13 in the Jones *et al*. (2003) map would correspond to the homoeologous pair B in the Barrett *et al*. (2004) map. This was confirmed in another study by Jones (2005), where the same two markers from Jones *et al*. (2003) were assigned to the same linkage group.

The level of segregation distortion in the two parental maps was very low, with 5.35% of the markers deviating from expected Mendelian ratios in R3R4 and 1% of the markers distorted in S1S4 (P values < 0.05).

Table 2.17. This table shows the number of mapped molecular markers and the linkage analysis of the R3R4 and S1S4 parental maps.

Parameters	R3R4	S1S4
AFLP		
Scored	233	291
Mapped	167 (72)	220 (76)
Unlinked	66 (28)	71 (24)
SSRs		
Scored	86	101
Mapped	55 (64)	81 (80)
Unlinked	31 (36)	20 (20)
Total		
Scored	319	392
Mapped	222 (70)	301 (77)
Unlinked	97 (30)	91 (23)
Linkage analysis		
Linkage groups	15	20
Map length (cM)	1138	1494
Map density (cM/marker)	5.13	4.96

Table 2.18. Distribution of AFLP and the five sets of SSR marker alleles in different linkage groups of R3R4 parental map.

Linkage groups	Origin of marker alleles							Map length (cM)	Map density (cM/marker)
	AFLP	WCSSR[a]	G-SSR[b]	TRSSR[c]	BES SSR[d]	BS SSR[e]	Total		
R-1(A)[*]	9	2	1	0	1	0	**13**	56	4.3
R-2(A)	9	1	0	0	0	0	**10**	65	6.5
R-3(A)	15	3	0	0	2	0	**20**	68	3.4
R-4(B)	13	2	0	2	0	0	**17**	84	4.9
R-5(C)	15	0	0	0	2	4	**21**	109	5.2
R-6(C)	10	3	0	0	0	1	**14**	128	9.1
R-7(D)	15	5	0	1	0	0	**21**	118	5.6
R-8(D)	25	3	0	0	2	0	**30**	114	3.8
R-9(E)	10	1	0	0	2	0	**12**	75	6.2
R-10(E)	11	2	0	0	0	0	**13**	89	6.8
R-11(F)	4	2	2	0	1	0	**9**	40	4.4
R-12(F)	5	2	2	0	0	0	**9**	36	4.0
R-13(G)	11	2	0	0	0	0	**13**	82	6.3
R-14(?)[**]	10	0	0	0	4	0	**14**	37	2.6
R-15(?)	5	0	0	0	0	0	**5**	37	7.4
Total	**167**	**28**	**5**	**3**	**14**	**5**	**222**	**1138**	**5.1**

[a] White clover SSR from Barrett *et al*. (2004)
[b] Genbank SSRs
[c] White clover SSRs from Jones *et al*. (2003)
[d] SSRs from the BAC-end sequence analysis (Section 3.2.6)
[e] SSRs from the complete BAC clone sequencing (Section 4.2.3)
[*] The number in brakets represent the nomenclature of linkage groups from Barrett *et al*. (2004)
[**] ? = the linkage groups have not yet been identified according to the previous nomenclature

Table 2.19. Distribution of AFLP and the five sets of SSR marker alleles in different linkage groups of S1S4 parental map.

Linkage groups	Origin of marker alleles							Map length (cM)	Map density (cM/marker)
	AFLP	WCSSR[a]	G-SSR[b]	TRSSR[c]	BES SSR[d]	BS SSR[e]	Total		
S-1(A)[*]	9	2	0	0	0	0	**11**	73	6.6
S-2(A)	5	0	0	0	1	0	**6**	50	8.3
S-3(B)	13	5	2	2	1	0	**23**	85	3.7
S-4(B)	22	4	0	1	0	0	**27**	94	3.5
S-5(C)	11	1	0	0	0	4	**16**	101	6.3
S-6(C)	4	2	0	0	0	2	**8**	44	5.5
S-7(D)	9	1	0	0	1	0	**11**	22	2.0
S-8(D)	24	1	0	2	2	1	**30**	111	3.7
S-9(D)	10	2	0	1	0	2	**15**	50	3.3
S-10(D)	6	0	0	0	1	0	**7**	18	2.6
S-11(E)	18	4	0	0	0	0	**22**	164	7.4
S-12(E)	12	4	0	0	0	1	**17**	133	7.8
S-13(F)	11	3	4	0	3	0	**21**	90	4.3
S-14(G)	10	4	0	0	0	1	**15**	90	6.0
S-15(G)	13	3	0	0	4	0	**20**	89	4.4
S-16(H)	7	3	0	1	0	0	**11**	50	4.5
S-17(H)	12	3	0	0	0	0	**15**	93	6.2
S-18(H)	7	2	0	0	0	0	**9**	31	3.4
S-19(?)[**]	5	0	0	0	0	0	**5**	34	6.8
S-20(?)	12	0	0	0	0	0	**12**	72	6.0
Total	**220**	**44**	**6**	**7**	**13**	**11**	**301**	**1494**	**4.9**

[a] White clover SSR from Barrett *et al*. (2004)
[b] Genbank SSRs
[c] White clover SSRs from Jones *et al*. (2003)
[d] SSRs from the BAC-end sequence analysis (Section 3.2.6)
[e] SSRs from the complete BAC clone sequencing (Section 4.2.3)
[*] The number in brakets represent the nomenclature of linkage groups from Barrett *et al*. (2004)
[**] ? = the linkage groups have not yet been identified according to the previous nomenclature

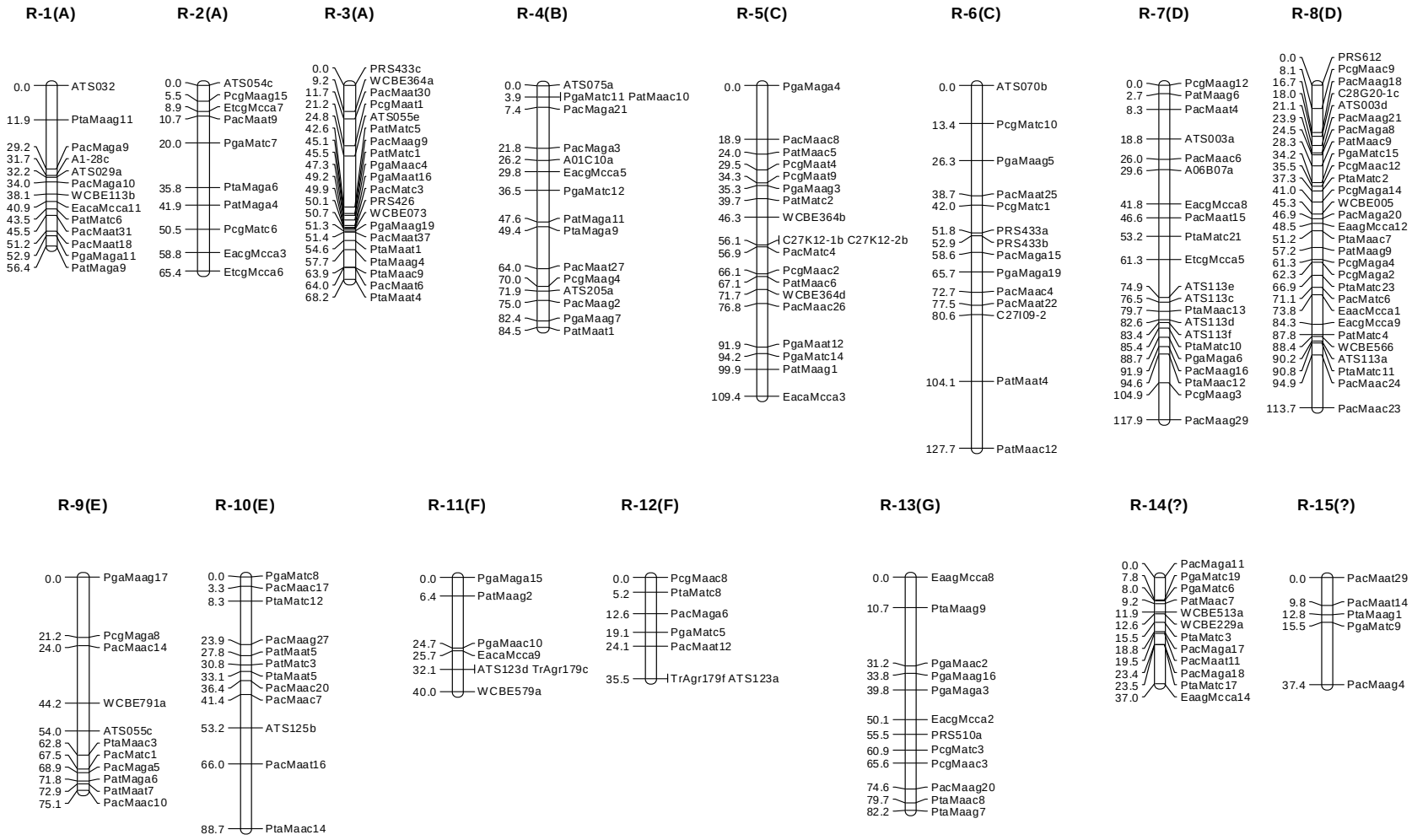


Figure 2.6. A genetic linkage map of parental line R3R4 based on AFLP and SSR markers. Map distances and marker names are shown on the left and right sides of the linkage groups. The nomenclature and orientation of the linkage groups (A to H) is given according to previously published white clover linkage map (Barrett *et al.*, 2004), with the exception of the number of homoeologues within each pair. The last two linkage groups, named "?", are as yet uncharacterised according to the previous nomenclature.

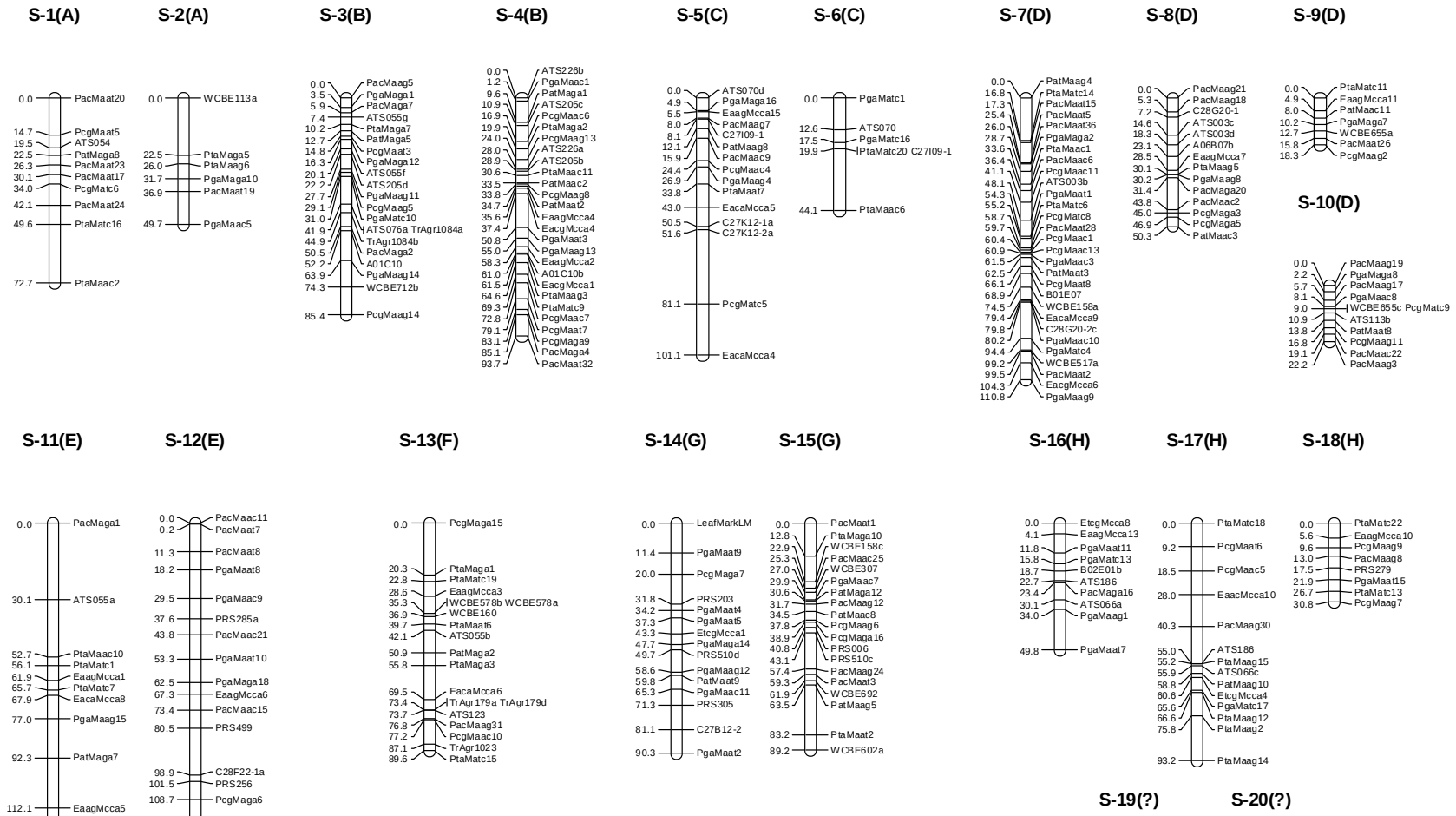


Figure 2.7. A genetic linkage map of parental line S1S4 based on AFLP and SSR markers. Map distances and marker names are shown on the left and right sides of the linkage groups. The nomenclature and orientation of the linkage groups (A to H) is given according to previously published white clover linkage map (Barrett *et al.*, 2004), with the exception of the number of homoeologues within each pair. The last two linkage groups, named "?", are as yet uncharacterised according to the previous nomenclature.

2.3.4. Morphological measurements

2.3.4.1. Glasshouse measurements

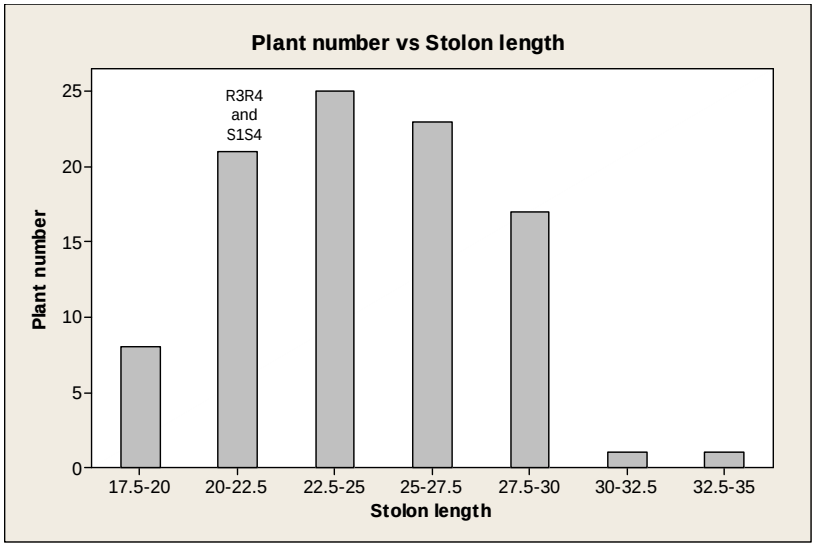
Glasshouse measurements were carried out as described in Section 2.2.6.1 on the mapping parents R3R4 and S1S4 (Figure 2.8) and the 94 progeny. The measurements of morphological traits exhibited apparently normal distributions (Figure 2.9) with substantial variation (Table 2.20).

Figure 2.8. Photo of the mapping parents R3R4 and S1S4 as grown in the glasshouse.

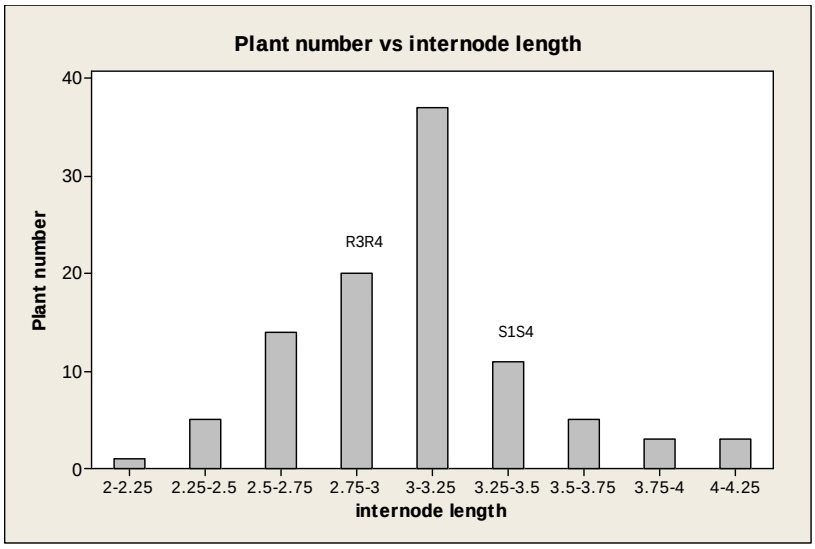
Table 2.20. Mean, standard deviation (SD) and coefficient of variation (CV) of morphological traits in the mapping parents and the mapping family for the glasshouse experiment.

	Family			R3R4	S1S4
	Mean	SD	CV	Mean	Mean
Stolon length (cm)	24.44	3.15	12.89	22.00	22.25
Internode length (cm)	3.11	0.43	13.93	2.75	3.38
Internode diameter (cm)	1.78	0.14	7.80	2.00	1.73
Petiole length (cm)	12.76	2.39	18.67	10.90	12.70
Leaf length (cm)	1.72	0.15	9.18	1.50	1.80
Leaf width (cm)	1.47	0.12	8.40	1.50	1.40

A



B



C

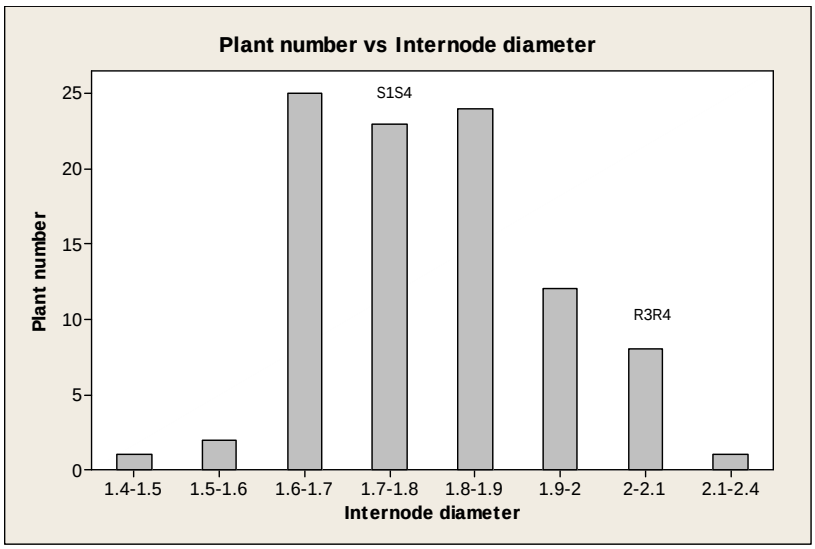
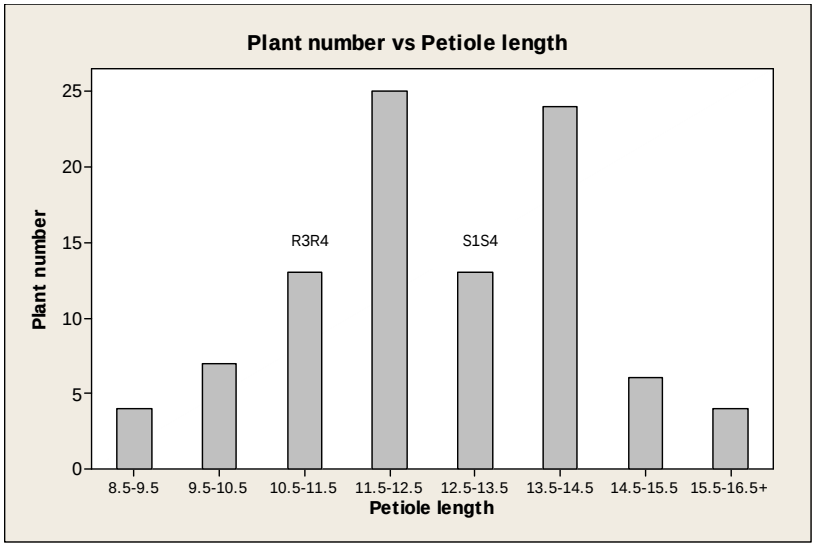
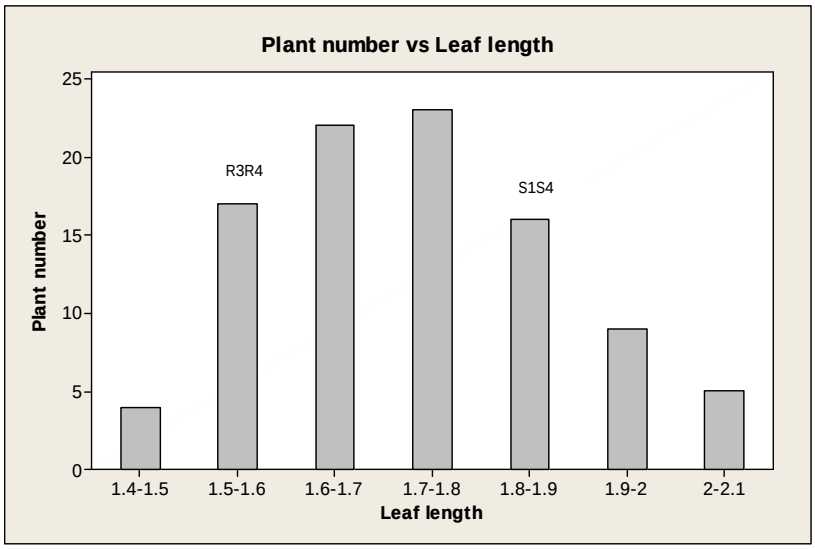


Figure 2.9: Frequency distribution of classes for morphological traits measured in the mapping population measured in the glasshouse. The classes in each figure are based on the mean values of four replicates per individual genotype in the mapping family growing in the glasshouse. A = Stolon length, B = Internode length, C = Internode diameter, D = Petiole length, E = Leaf length, F = Leaf width.

D



E



F

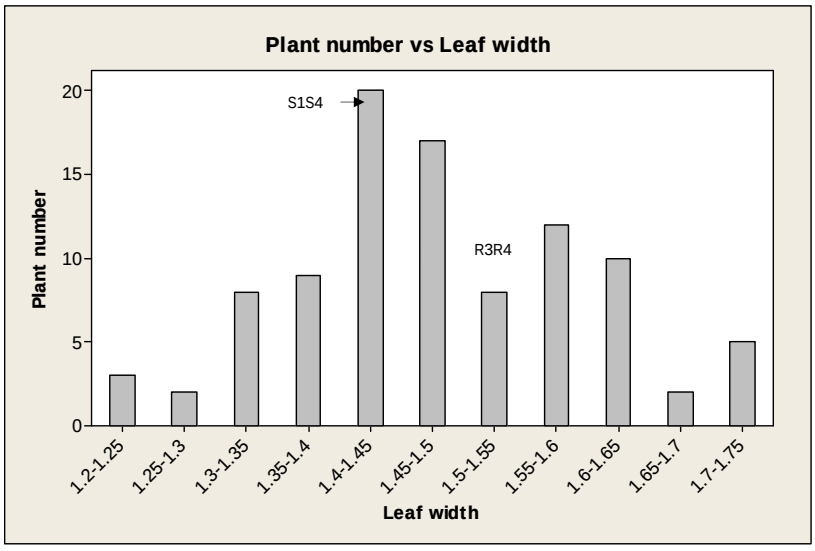


Figure 2.9. Continued.

Stolon length ranged from 17.5 – 33.15 cm with a mean value for all plants of 24.44 cm. Parental values were 22.25 cm for R3R4 and 22 cm for S1S4. Internode length ranged from 2.18 – 4.18 cm with a mean value of 3.11 cm. The R3R4 parent had a mean length of 3.38 cm and the S1S4 parent 2.75 cm. Internode diameter ranged from 1.45 – 2.35 cm and the mean of all plants was 1.78 cm. R3R4 had an internode diameter of 1.72 cm while S1S4 was 2 cm. Petiole length values ranged from 8.78 – 29.5 cm with a mean of 12.76 cm. Parental values were 12.65 cm (R3R4) and 10.96 cm (S1S4). Leaf length ranged from 1.4 – 2.1 cm and the mean for all plants was 1.7 cm. R3R4 had a leaf length of 1.75 cm and S1S4 of 1.5 cm. Leaf width ranged from 1.2 – 1.75 cm with a mean of 1.47 cm. Parental values were 1.35 cm (R3R4) and 1.52 cm (S1S4).

Many of the traits being examined in the glasshouse (and later the field) experiment are likely to be interdependent, and have a common genetic and developmental basis. In order to identify the occurrence of this phenomenon for the traits analysed, correlation coefficients between the traits were calculated using the replicate means (Table 2.21). There were a number of highly significant positive correlations, for example between stolon length and internode length (0.701***), between leaf length and leaf width (0.730***) and also between the leaf morphology (length and width) and all other traits. Scatter plots with regression lines were drawn to demonstrate the correlation between each trait (Figure 2.10).

Table 2.21. Table of correlation coefficients from traits measured in the mapping family in the glasshouse.

	Stolon length	Internode length	Internode diameter	Petiole length	Leaf length
Internode length	0.701***				
Internode diameter	0.227*	0.325*			
Petiole length	0.234*	0.216*	0.139		
Leaf length	0.450***	0.459***	0.441***	0.408***	
Leaf width	0.391***	0.350***	0.458***	0.500***	0.730***

*, *** show significance at $p < 0.1$, $p < 0.001$.

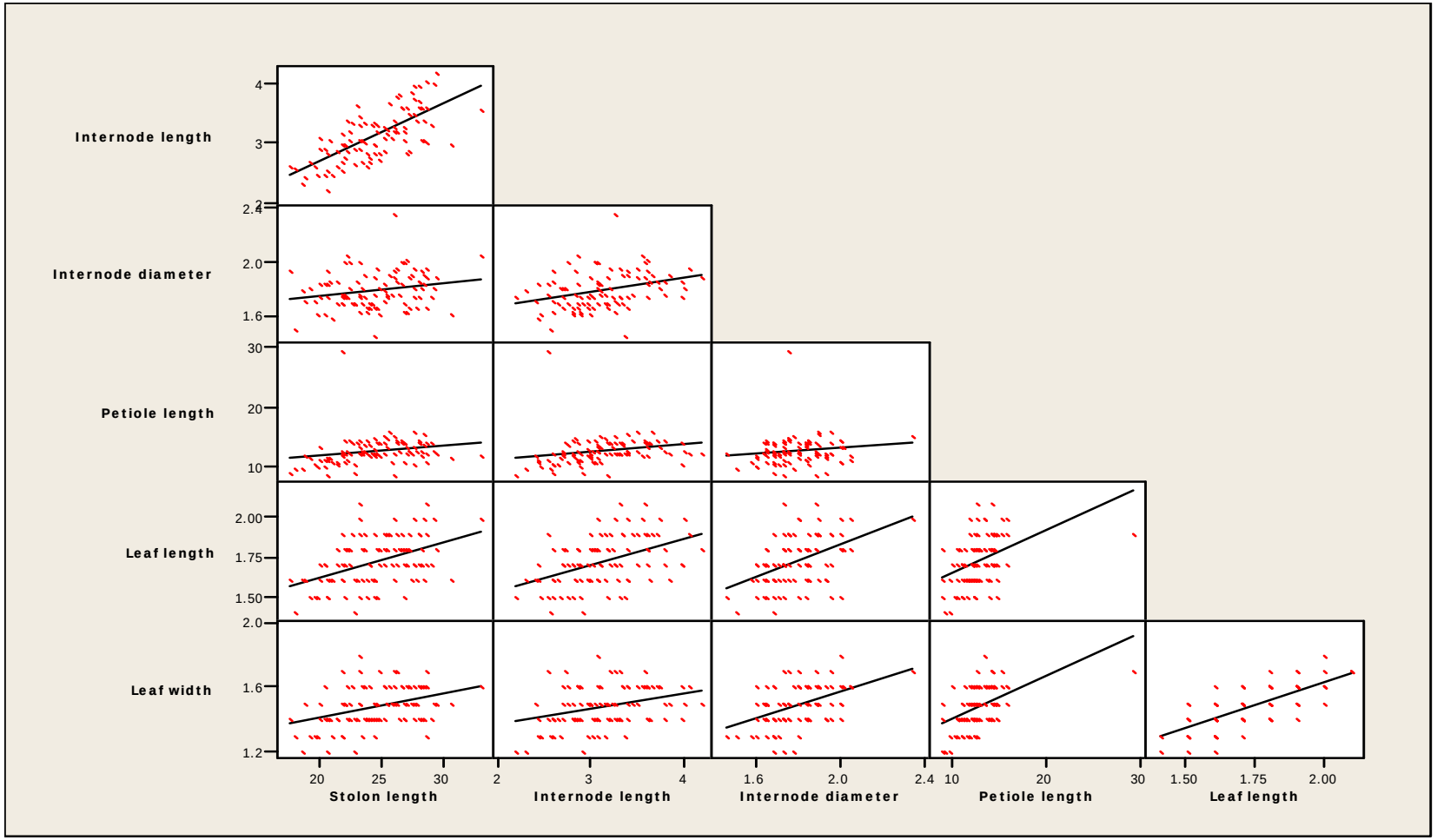


Figure 2.10: Scatter plots of traits measured in the glasshouse

2.3.4.2. Field measurements

Field measurements were carried out as described in Section 2.2.6.2 on the mapping parents (Figure 2.11 and 2.12) and the progeny. For each trait measured the frequency of distribution was plotted displaying apparently normal distribution (Figure 2.13) with considerable variation (Table 2.22).

Figure 2.11. Photo of the parental plant S1S4 as grown in the field (Taken on the same date as R3R4).

Figure 2.12. Photo of the parental plant R3R4 as grown in the field (Taken on the same date as S1S4).

Table 2.22. Mean, standard deviation (SD) and coefficient of variation (CV) of morphological traits in the mapping parents and the mapping family for the field experiment.

	Family			R3R4	S1S4
	Mean	SD	CV	Mean	Mean
Flowering time	49.20	2.96	6.03	44.88	56.25
Leaf length (mm)	26.92	3.33	12.38	17.50	23.75
Leaf width (mm)	19.97	2.63	13.19	15.50	16.50
Internode length (mm)	32.13	6.63	20.63	21.00	37.50
Internode diameter (mm)	2.22	0.23	10.27	1.83	1.93
Petiole length (mm)	96.87	25.05	25.86	63.50	88.75
Plant spread (mm)	1076.68	108.22	10.05	1092.50	975.00
Plant height (mm)	201.27	32.70	16.25	160.00	167.50
Flower height (mm)	317.66	37.15	11.69	301.25	276.25

The flowering time ranged from 44 – 56.25 days with a mean value for all plants of 49.20 days. Parental flowering times were 44.88 days (R3R4) and 56.25 days (S1S4). Leaf length ranged from 17.50 – 34.50 mm and the mean was 26.92 mm. Parental means were 17.50 and 23.75 mm for R3R4 and S1S4. Leaf width ranged from 15 – 26.75 mm with a mean of 19.97 mm. R3R4 had a leaf width of 15.50 mm and S1S4 of 23.75 mm. Internode length ranged from 16.75 – 50.25 mm with the mean of all plants was 32.13 mm, with R3R4 and S1S4 parents 21 mm and 37.50 mm respectively. Internode diameter ranged from 1.63 – 2.75 mm and a mean of 2.22 mm. The means for the parents were 1.83 mm (R3R4) and 1.93 mm (S1S4). Petiole length ranged from 46.5 – 165.75 mm with a mean of all plants of 96.87 mm. Parental values were 63.50 mm for R3R4 and 88.75 for S1S4.

Stolon length was not measured in the field due to the volume of the plant growth compared the glasshouse. In addition, plant spread and height and flower height were measured in the field. Values for plant spread ranged from 800 – 1385 mm with a mean of 1076.68 mm. Parental means were 1092.50 mm (R3R4) and 975 mm (S1S4). Plant height ranged from 132.50 – 263.75 mm and the mean of all plants was 201.27 mm. The parent R3R4 had a plant height of 160 mm and 167.50 for S1S4. Mean flower height ranged from 237.50 – 416.25 mm with an average of 317.66 mm. Parental values were 301.25 mm and 276.25 mm for R3R4 and S1S4 respectively.

As in the glasshouse data, the possibility exists for significant correlation between many of the morphological traits measured on the field. In this case, in addition to

possible common developmental and genetic interdependence, some of the traits measured are very likely to be components contributing to variation in other traits (for instance stolon and internode length are likely components of traits such as plant spread and plant height). As expected, significant ($p<0.001$) positive correlation coefficients between many of the traits were seen (Table 2.23, Figure 2.14).

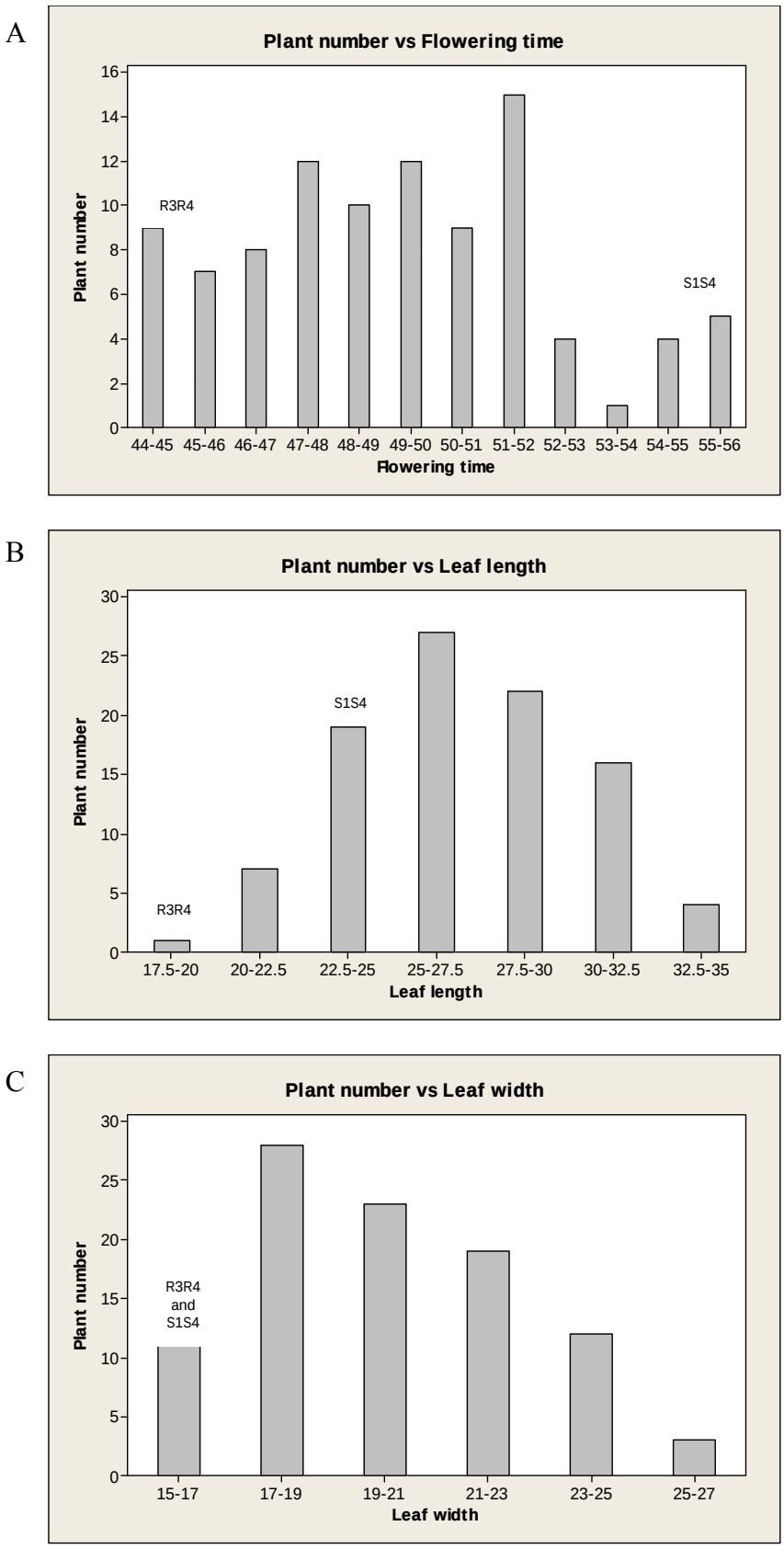
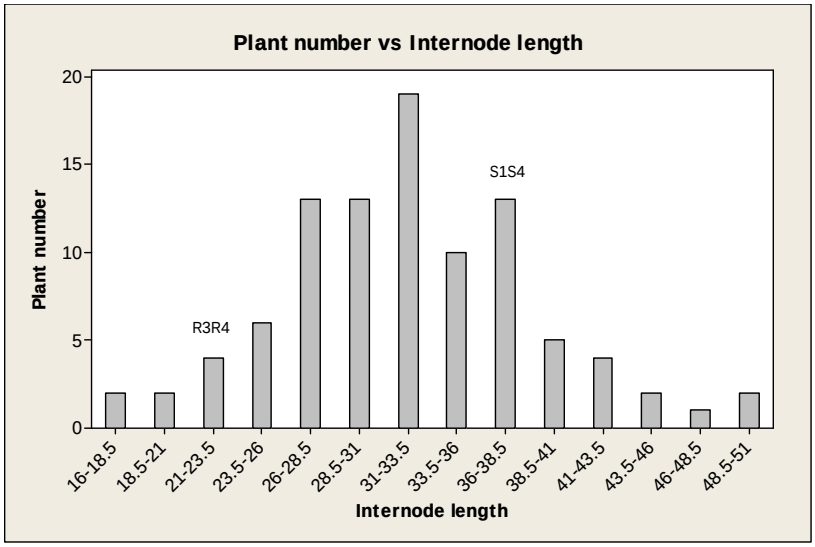
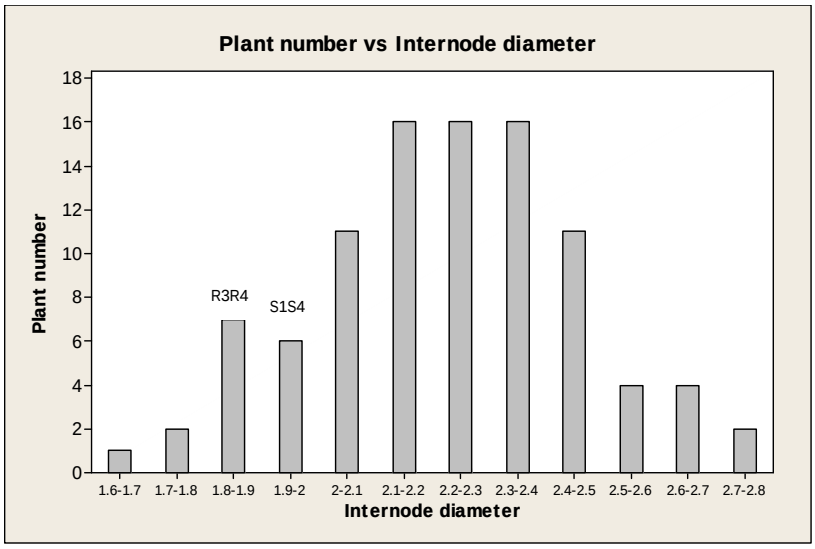


Figure 2.13: Frequency distribution of classes for morphological traits measured in the mapping population measured in the field. The classes in each figure are based on the mean values of four replicates per individual genotype in the mapping family growing in the field. A = Flowering time, B = Leaf length, C = Leaf width, D = Internode length, E = Internode diameter, F = Petiole length, G = Plant spread, H = Plant height, I = Flower height.

D



E



F

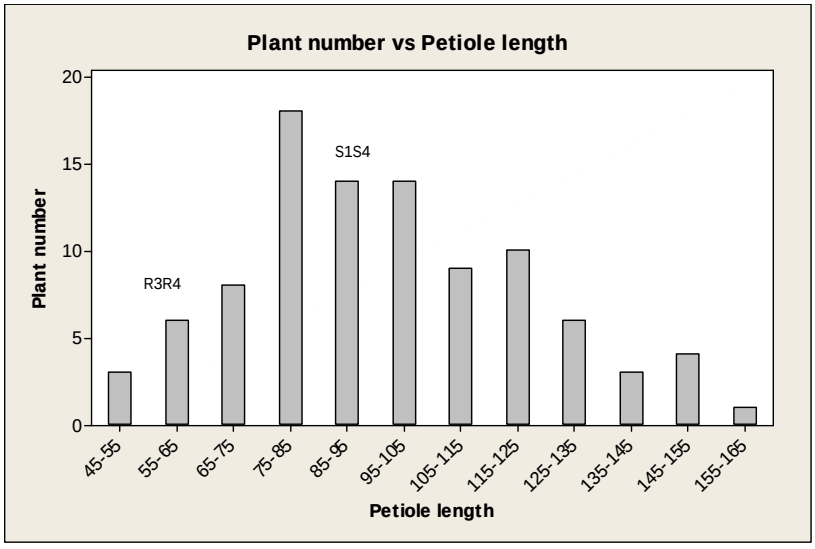
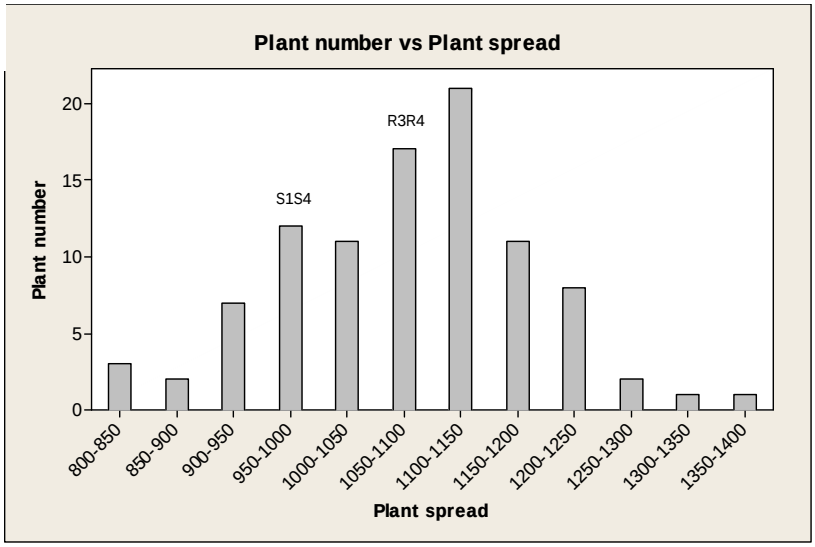
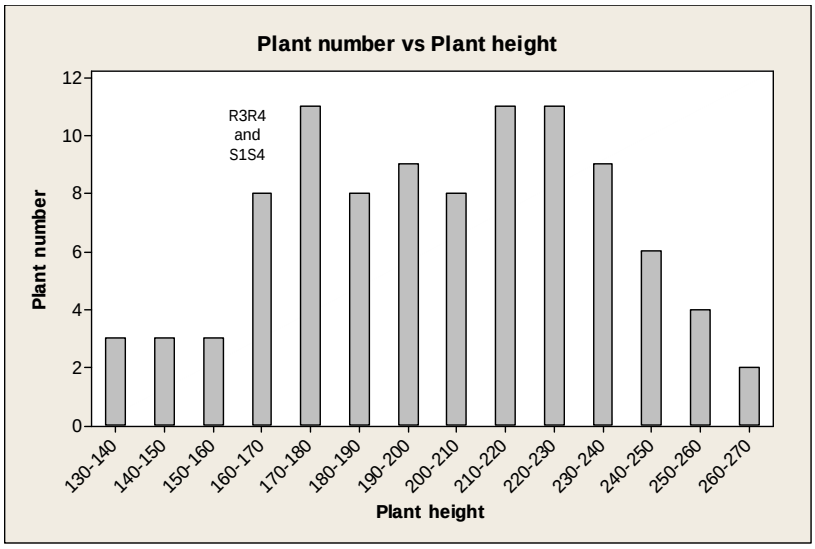


Figure 2.13. Continued.

G



H



I

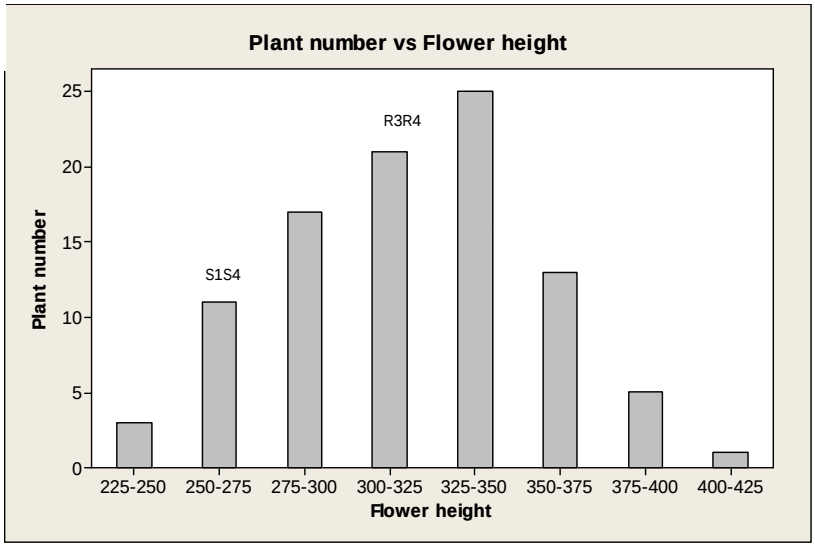


Figure 2.13. Continued.

Table 2.23. Table of correlation coefficients from traits measured in the mapping family in the field.

	Flowering time	Leaf length	Leaf width	Internode length	Internode diameter	Petiole length	Plant spread	Plant height
Leaf length	0.140							
Leaf width	0.082	0.817***						
Internode length	0.303*	0.517***	0.521***					
Internode diameter	0.020	0.644***	0.686***	0.489***				
Petiole length	0.183*	0.656***	0.622***	0.496***	0.427***			
Plant spread	-0.053	0.497***	0.350***	0.391***	0.488***	0.348***		
Plant height	0.137	0.600***	0.575***	0.331***	0.537***	0.621***	0.518***	
Flower height	-0.062	0.407***	0.398***	0.257*	0.392***	0.468***	0.536***	0.718***

*, *** show significance at $p < 0.1$, $p < 0.001$.

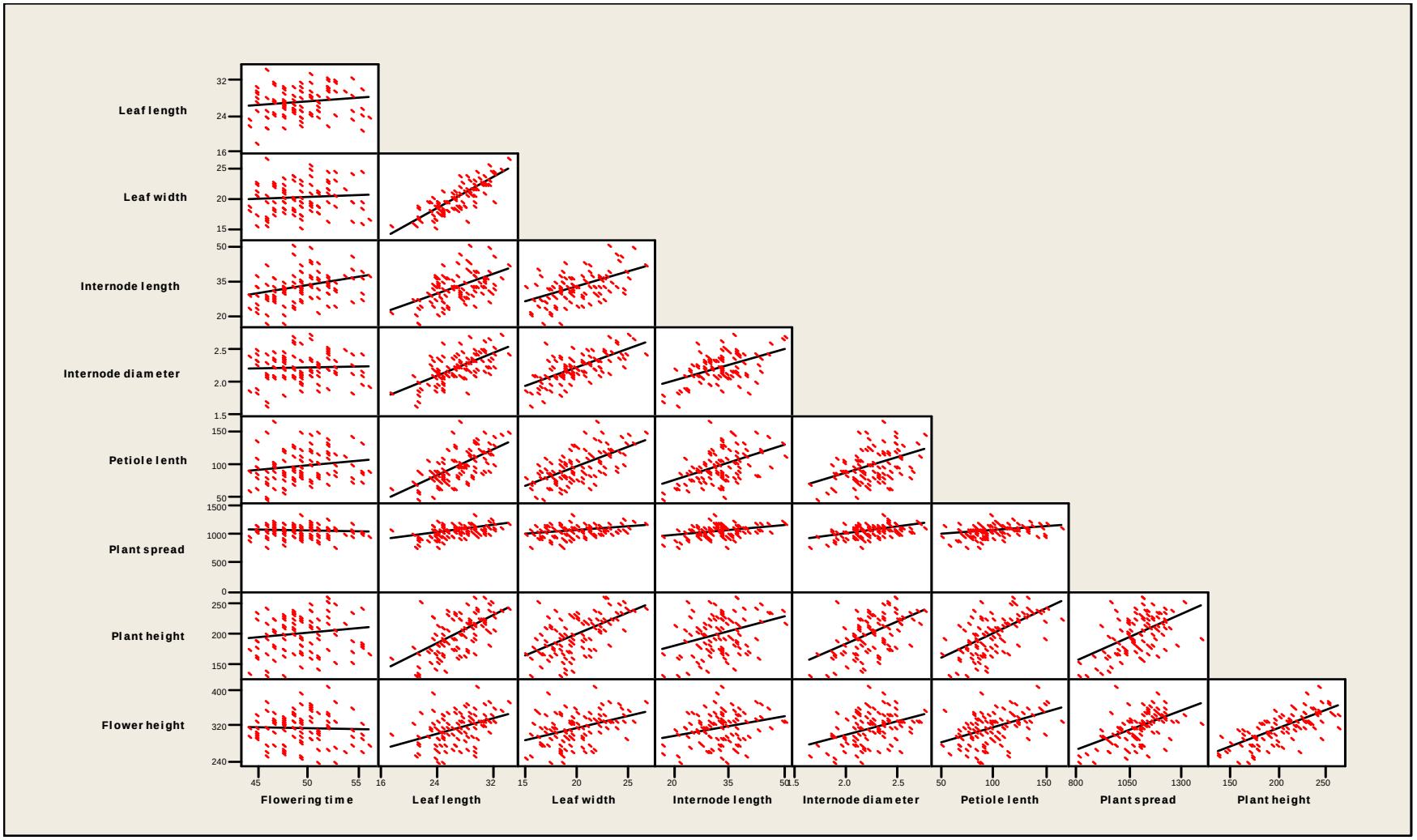


Figure 2.14. Scatter plots of traits measured in the field.

2.3.4.3 Correlation between field and glasshouse data

Several traits (leaf length and width, petiole length and internode length and diameter) were examined in both the field and glasshouse studies, introducing an element of replication across environments into the study. In order to examine the effect of differing environments on the traits, scatterplots (Figure 2.15) were generated for each of the traits in the respective environments, and correlation coefficients between the traits in both environments were calculated (Table 2.24). When the two environments were compared, the highest correlations were observed for internode diameter, leaf width and leaf length, with correlation coefficient of 0.425, 0.400 and 0.399 respectively (Table 2.25). In the case of petiole length, the scatterplot looked as if it displayed the highest correlation (Figure 2.15), but the correlation coefficient was only of 0.295. This could be explained by a conflict between the two environments.

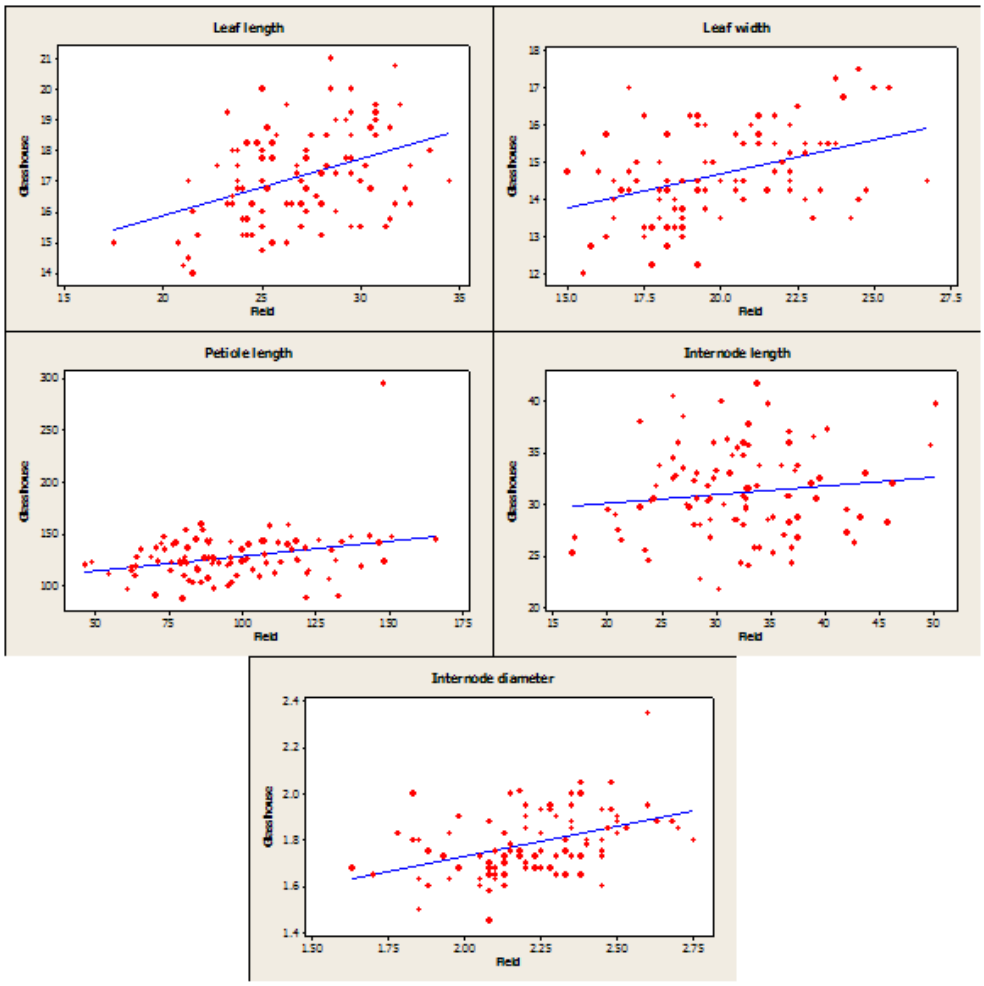


Figure 2.15. Scatterplots of each common trait between the glasshouse experiment and the field experiment.

Table 2.24. Coefficient of correlation of common traits between the glasshouse experiment and the field experiment.

Trait	Coefficient of correlation
Leaf length	0.399***
Leaf width	0.400***
Internode length	0.130*
Internode diameter	0.425***
Petiole length	0.295*

*, *** show significance at $p < 0.1$, $p < 0.001$.

As expected, in absolute terms there were considerable differences in growth between plants in the two environments. Plants in the field were in a nursery with 90 cm between each plant, which represents an ideal growth condition. However, plants in the glasshouse were grown in relatively small pots, which considerably limited the plant development. Plant morphological measurements were higher in the field in comparison with the glasshouse results, except for the petiole length where the average length was higher in the glasshouse. This could be explained by the fact that the plants were restricted in small pots and therefore stolon length and internode development were reduced and those were compensated by a better elongation of the petiole.

2.3.5. Quantitative trait analysis

2.3.5.1. Field measurements analysis

Significant associations between marker and trait data were established for all traits in each parental map as summarised in Table 2.25 and Figure 2.16 for S1S4 and in Table 2.26 and Figure 2.17 for R3R4. A total of 23 QTLs for S1S4 and 11 QTLs for R3R4 were identified using simple interval mapping (SIM) with a LOD cut-off of 2.5 and 1 to 4 QTLs were detected for each trait in S1S4 and 0 to 2 in R3R4. The phenotypic variation explained by each QTL (r^2) ranged from 3.1 to 16.6% in S1S4 and 5.4 to 20.9% in R3R4.

(i) Flowering time. Four QTLs were identified for flowering time in S1S4 on linkage groups (LGs) S-1(A), S-4(B), S-11(E) and S-12(E), and the phenotypic variation explained by each QTL ranged from 3.1 to 13.7%. Collectively these four QTLs explained 34.1% of the phenotypic variation. Similarly, two QTLs were identified in R3R4 on LGs R-5(C) and R-10(E). The total phenotypic variation explained these QTLs was 21.85%. It is interesting to note that QTLs for flowering time were found

on both homoeologues of LG E of the two in parent S1S4 and one of the two homoeologues of parent R3R4. The QTLs all mapped to a similar position across the 3 LGs.

(ii) Leaf length. One QTL was identified in S1S4 at the top of LG S-4(B) explaining 12.6% of the phenotypic variation. One QTL was also detected in R3R4 at the bottom of LG R-4(B), explaining 8% of the phenotypic variation. Even though the two QTLs were both found on LG B of the two parents, their relative positions did not seem to correspond to each other.

(iii) Leaf width. Three QTLs were detected in S1S4 on linkage groups S-1(A), S-4(B) and S-14(G). The phenotypic variation explained by the loci totalled 29.5%. Likewise, two QTLs were identified in R3R4 on linkage groups R-4(B) and R-8(D) explaining a total of 18% of the phenotypic variation.

(iv) Internode length. One QTL was detected on the linkage group S-4(B) of S1S4 accounting for 9.8% of the phenotypic variation for the trait, and one QTL was also detected on LG R-8(D) of R3R4 (accounting for 14.3% of the phenotypic variation).

(v) Internode diameter. Four QTLs were found in S1S4 on LGs S-4(B), S-9(D), S-14(G) and S-15(G), collectively accounting for 40.3% of the phenotypic variation. No QTL was detected in R3R4. The two QTLs on LG G of parent S1S4 were found on approximately the same position on the two homoeologues.

(vi) Petiole length. Two QTLs were identified in S1S4 on linkage groups S-12(E) and S-15(G). The total phenotypic variation explained by these loci was 20.75%. In R3R4, one QTL was detected on LG R-8(D). This was the single most robust QTL in the field experiment, with a LOD score of 6.12, and accounting for 19.8% of the phenotypic variation for the trait.

(vii) Plant spread. One QTL was found in S1S4 on LG S-12(E) (explaining 5.8% of the phenotypic variation) and one QTL was identified in R3R4 on LG R-8(G) explaining 6.8% of the phenotypic variation).

(viii) Plant height. Four QTLs were detected in linkage groups S-4(B), S-7(D), S-12(E) and S-15(G) of the S1S4 map. The phenotypic variation explained by these loci totalled 34.9%. In the parental map R3R4, two QTLs were found on linkage groups R-4(B) and R-8(D). The cumulative phenotypic variation explained by these two QTLs was 26.3%**.**

(ix) Flower height. Three QTLs were identified in S1S4 on LGs S-12(E), S-15(G) and S-20(?) explaining a total of 33.7% of the phenotypic variation for the trait. In R3R4, only one QTL was detected on LG R-8(D), explaining 12.9% of the phenotypic variation.

Table 2.25. Summary of QTL detection information for the nine phenotypic traits measured in the field for the S1S4 parental map.

		S1S4			
		Kruskal-Wallis		SIM	
Trait	LG	Position (cM) P<0.01	Max r^2 (%)	Max LOD	Position (cM)
Flowering time	S-1(A)	-	3.09	5.06	19.5
	S-4(B)	19.93 – 37.39	6.9	2.62	24.0
	S-11(E)	0 – 67.85	13.66	2.79	56.1
	S-12(E)	43.76 – 67.25	10.36	3.43	62.5
Leaf length	S-4(B)	0 – 1.19	12.58	2.72	1.19
Leaf width	S-1(A)	-	6.22	3.82	19.5
	S-4(B)	27.0 – 29.0	15.09	3.69	28.9
	S-14(G)	0 – 65.25	8.21	2.56	34.2
Internode length	S-4(B)	68.0 – 70.0	9.76	2.91	69.3
Internode diameter	S-4(B)	9.56 – 37.93	16.58	3.72	28.9
	S-9(D)	4.93 – 12.72	9.54	2.72	10.2
	S-14(G)	0 – 65.25	10.38	2.64	25.0
	S-15(G)	-	3.84	2.56	37.8
Petiole length	S-12(E)	53.27 – 80.51	8.95	2.93	72.3
	S-15(G)	34.47 – 43.09	11.83	2.76	38.9
Plant spread	S-12(E)	98.9 – 133.45	5.83	2.56	118.7
Plant height	S-4(B)	27.0 – 29.0	5.48	5.25	28.9
	S-7(D)	94.44 – 110.75	7.7	3.26	94.4
	S-12(E)	37.63 – 80.51	8.31	4.02	48.8
	S-15(G)	38.89 – 43.09	13.42	3.28	38.9
Flower height	S-12(E)	53.27 – 133.48	10.61	3.33	118.7
	S-15(G)	26.99 – 43.09	10.88	3.08	38.9
	S-20(?)	0 – 13.03	12.26	2.92	13.0

Table 2.26. Summary of QTL detection information for the nine phenotypic traits measured in the field for the R3R4 parental map.

R3R4					
		Kruskal-Wallis		SIM	
Trait	LG	Position (cM) P<0.01	Max r^2 (%)	Max LOD	Position (cM)
Flowering time	R-5(C)	91.94 – 109.39	15.85	4.20	99.2
	R-10(E)	3.28 – 23.89	6.02	4.19	18.3
Leaf length	R-4(B)	-	8.01	3.85	71.9
Leaf width	R-4(B)	-	6.79	3.96	71.9
	R-8(D)	0 – 45.32	11.18	3.70	45.3
Internode length	R-8(D)	0 – 45.32	14.33	3.71	45.3
Petiole length	R-8(D)	0 – 62.23	19.82	6.12	35.5
Plant spread	R-8(D)	34.16 – 57.19	6.81	2.81	41.0
Plant height	R-4(B)	-	5.43	3.39	71.9
	R-8(D)	0 – 62.23	20.92	5.85	56.9
Flower height	R-8(D)	0 – 62.23	12.89	3.49	34.2

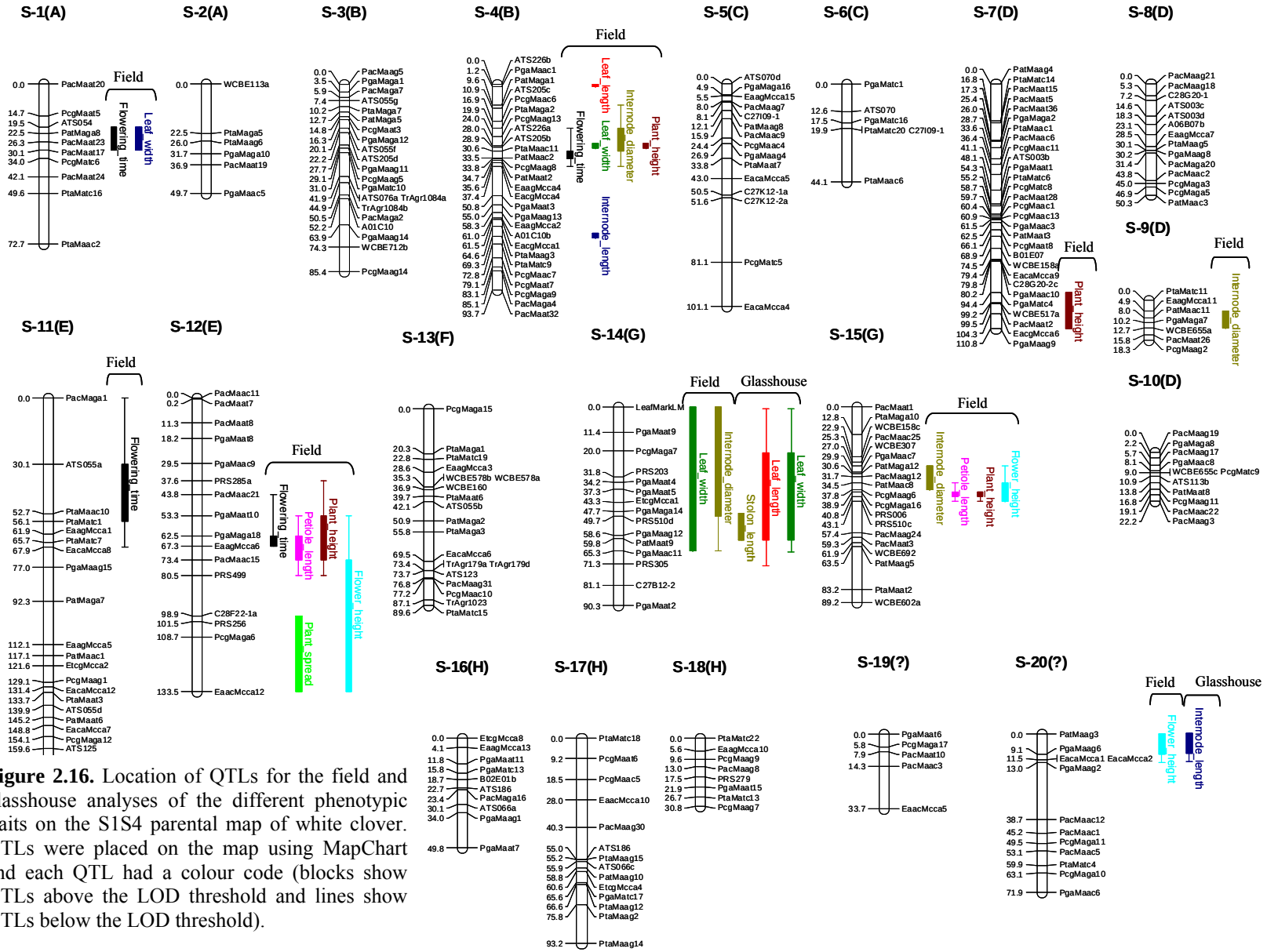


Figure 2.16. Location of QTLs for the field and glasshouse analyses of the different phenotypic traits on the S1S4 parental map of white clover. QTLs were placed on the map using MapChart and each QTL had a colour code (blocks show QTLs above the LOD threshold and lines show QTLs below the LOD threshold).

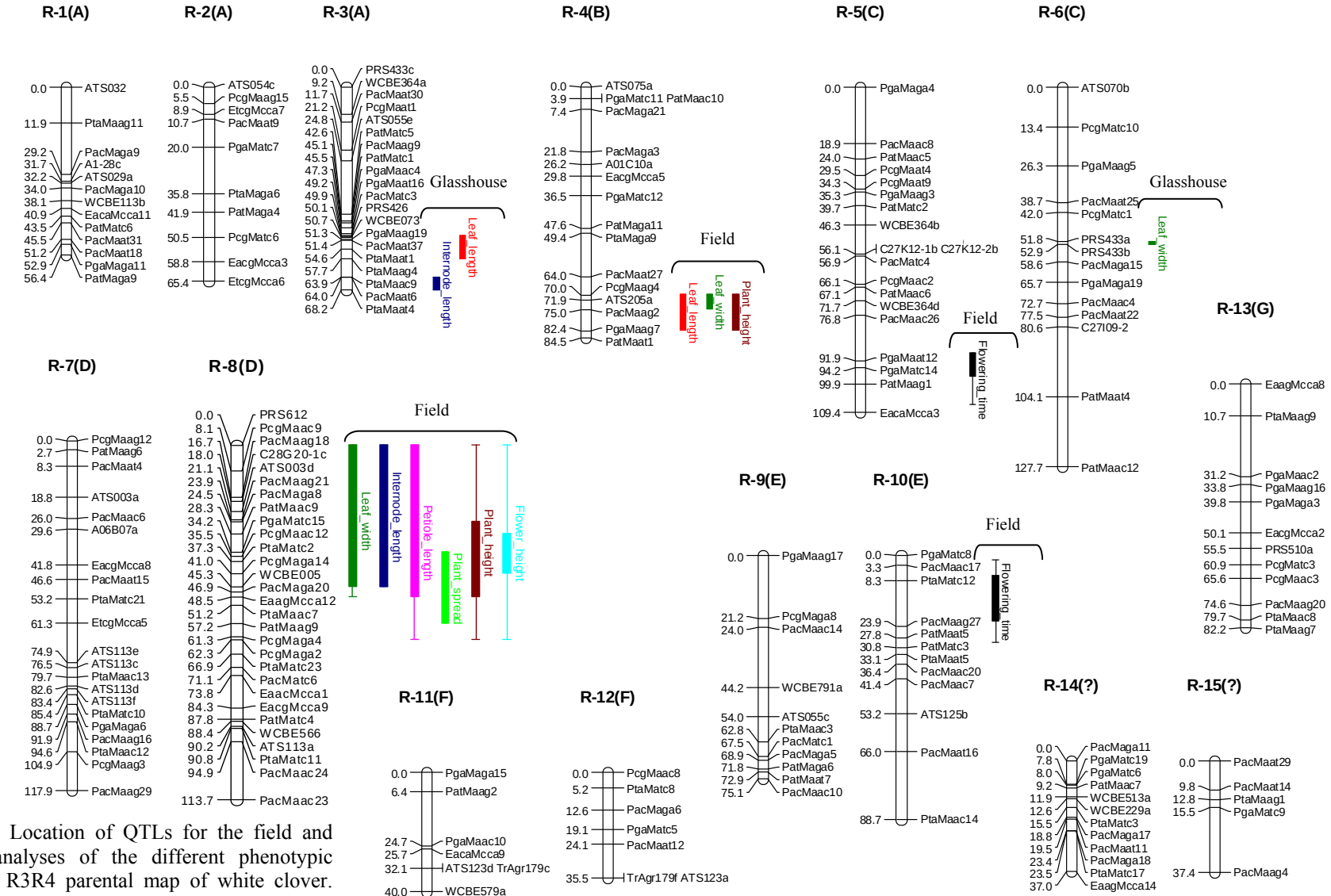


Figure 2.17. Location of QTLs for the field and glasshouse analyses of the different phenotypic traits on the R3R4 parental map of white clover. QTLs were placed on the map using MapChart and each QTL had a colour code (blocks show QTLs above the LOD threshold and lines show QTLs below the LOD threshold).

2.3.5.2. Glasshouse measurements analysis

Significant QTLs were identified for the 6 measurements carried out in the glasshouse as summarised in Table 2.27 and Figure 2.16 for S1S4 and Figure 2.17 for R3R4. A total of 3 QTLs were detected in the parental map R3R4 with a phenotypic variation explained by each QTL ranging from 7.3% to 10.3%. For the parental map S1S4, 4 QTLs were identified. The phenotypic variation explained by each QTL ranged from 6.3% to 15.6%.

(i) Stolon length. One QTL for stolon length was identified on linkage group S-14(G) for S1S4 (explaining 6.3% of the phenotypic variation).

(ii) Internode length. Similarly to stolon length, only one QTL was found in each map. For R3R4, the QTL was on linkage group R-3(A) and displayed a phenotypic variation of 7.8%. The QTL in S1S4 was located on linkage group S-20(?), and the phenotypic variation explained was 14.3%.

(iii) Internode diameter. No QTL for internode diameter was detected in either R3R4 or S1S4.

(iv) Petiole length. No QTL for petiole length was detected in either R3R4 or S1S4.

(v) Leaf length. One QTL was detected for leaf length for both parental maps: the QTL in R3R4 was found on linkage group R-3(A), explaining 10.3% of the phenotypic variation; whereas the QTL in S1S4 was detected on linkage group S-14(G), explaining 15.6% of the phenotypic variation.

(vi) Leaf width. One QTL were identified in R3R4 on linkage group R-6(C) explaining 7.3% of the phenotypic variation. Only one QTL was detected in S1S4 on linkage group S-14(G). The phenotypic variation explained by this QTL was 13.6%.

Table 2.27. Summary of QTL detection information for the six phenotypic traits measured in the glasshouse.

R3R4

Trait	LG	Kruskal-Wallis		SIM	
		Position (cM) P<0.01	Max r^2 (%)	Max LOD	Position (cM)
Internode length	R-3(A)	63.2 – 68.2	7.85	2.60	63.9
Leaf length	R-3(A)	49.9 – 57.7	10.34	2.86	50.7
Leaf width	R-6(C)	-	7.27	2.67	52.9

S1S4

Trait	LG	Kruskal-Wallis		SIM	
		Position (cM) P<0.01	Max r^2 (%)	Max LOD	Position (cM)
Stolon length	S-14(G)	-	6.33	2.96	47.7
Internode length	S-20(?)	0 – 13.0	14.28	3.52	11.5
Leaf length	S-14(G)	0 – 71.3	15.56	4.58	42.3
Leaf width	S-14(G)	0 – 65.3	13.63	3.92	59.8

2.4 Discussion

In this chapter, AFLP and SSR marker loci were used to construct a genetic linkage map of the allotetraploid white clover genome in the F_1 (R3R4 x S1S4) population. The parental maps were constructed largely with AFLPs, due to the ease with which a large number of markers can be readily generated to give substantial genome coverage with this technique. However, AFLPs are dominant, anonymous markers, making it difficult to align maps produced in different crosses using AFLPs alone. To overcome this problem, a number (59) of SSR markers were also used for mapping, in order to orient and identify the linkage groups generated during this study relative to previously generated maps in white clover. These SSR markers also served to identify homoeologue pairs within each parent. In this study, 25 SSR markers from Barrett *et al*. (2004) were scored dominantly in the mapping population giving rise to a total of 79 SSR alleles. The parental map R3R4 contained 28 of these SSR alleles, which helped us to identify 13 of the 15 linkage groups of this map, the other two linkage groups remained unidentified because no SSR was present in those LGs. The parental map S1S4 contained 44 SSR alleles from Barrett *et al*. (2004), which were distributed across 18 out of 20 linkage groups identified for this parent, the other two remaining unidentified.

The length of the genetic map of the male parent (R3R4) was 1138 cM and that of the female parent (S1S4) was 1494 cM. A similar length was observed in the microsatellite map of white clover from Barrett *et al*. (2004), which was 1134 cM; however the molecular based marker constructed by Jones *et al*. (2003) had a smaller length (825 cM). Both previously published maps were presented as single maps resulting from the merging of the two parental maps, in comparison with the individual parental maps presented here. This is almost certainly a reflection of the more extensive use of co-dominant markers, and/or a greater number of bi-parentally derived (3:1) AFLP markers in these previous maps, which allowed integration of the individual parental maps. The maps of S1S4 and R3R4 developed in this study were based largely on AFLP markers. The vast majority of the markers (85%) showed a uniparental (1:1) segregation pattern, while only 15% of the markers showed a biparental (3:1) segregation pattern. The very low number of bi-parentally-derived markers, which represent alleles shared by the two parents, indicates that the two parental genotypes were considerably different genetically, resulting in a highly

polymorphic mapping population. However as outlined by Maliepaard *et al*. (1998) dominant markers exhibiting a biparental (3:1) segregation ratio are useful for identifying homologous linkage groups between parental maps in outbreeding crosses. Almost none of the biparental markers that segregated in the population were mapped in both of the parental maps, making it impossible to use these markers to help integrate the two parental maps in JoinMap. It is arguable that an attempt could have been made to integrate the two maps on the basis of the SSR alleles mapped on both parents. However, two factors prevented this. Firstly, while the mapped SSRs had sufficient coverage to identify and orient many of the groups, it was felt that, overall, they gave too few anchor points to attempt map integration. Secondly, and perhaps more significantly, it was not possible to reliably identify homologous groups (e.g. A1 in one parent and A1 in the other parent) between the two maps (this is discussed further below), leading to the possibility of incorrectly assembling the incorrect sets of homologues between the parents (e.g. merging A1 of one parent with A2 of the other parent). Given these factors it was felt that it was more appropriate to represents the parental maps separately.

Although two genetic maps for white clover have so far been described (Jones *et al.*, 2003; Barrett *et al.*, 2004), limited marker transfer has occurred to align the maps, and the number of white clover genetic markers accessible in the public domain is still relatively small. The map presented in this study was constructed largely using AFLP markers, with a reasonable number of SSR markers from the map of Barrett *et al*., and the nomenclature and orientation of the parental maps was determined according to these markers (Barrett *et al.*, 2004). The latter arbitrary labelled the eight homoeologous pairs of linkage groups A to H, with subscript numerals to distinguish the homoeologue pairs. Unlike other, more characterised allopolyploids, such as hexaploid wheat, the subscripts do not designate the ancestral genomes of clover, but are also arbitrarily assigned. However, we did not attempt to distinguish between homoeologue pairs in the same way as Barrett *et al*. (2004), since it was apparent that the SSR markers, while consistently homoeologue group specific were not consistently "ancestral genome" specific. Recently, Ellison *et al*. (2006) studied the ancestral origin of white clover and identified two diploid *Trifolium* species, *T. occidentale* and *T. pallescens*, as putative ancestors for white clover. Those two putative parental species are from well-separated clades, which would lend hope to

the idea of, in the future, developing genetic markers that are consistently diagnostic of the ancestral origin of each homoeologue pair. The gene-associated SNPs recently described by Cogan *et al*. (2007) might represent a possible source of such genome specific markers.

Following the construction of the genetic linkage map, we carried out a preliminary investigation of association between molecular markers and a number of traits segregating in the mapping population. In this study, we focused on the analysis of quantitative traits related to plant morphology (leaf length and width, petiole length, internode length and diameter, plant height and spread) and flowering time. Even though the F_1 (R3R4 x S1S4) population was not specifically developed to exhibit morphological diversity, it was felt that sufficient diversity was observable to merit a QTL analysis in an attempt to contribute to the body of knowledge regarding the genetic control of these traits.

Morphological characteristics of plants are often heavily influenced by the environment. In these circumstances it is likely that different genetic loci will show significant associations with traits in different years and environments. Some of this apparent environmental and seasonal variation is due to data noise, while some may represent the involvement of different loci in a trait in response to different environmental conditions. In this study we have analysed several traits in two different environments, glasshouse and field, allowing us to compare the loci involved in the same trait across different environments. The expectation is that genetic intervals that exhibit a significant association with the same (or similar) traits across multiple environments represent quantitative trait loci which are less 'environmentally plastic'. In the future, these multi-environment, stable QTLs represent the best targets for strategies such as 'QTL-cloning' and marker-assisted selection.

Many of the traits examined in this study are highly correlated with each other, presumably because of a similar genetic and physiological basis, or because some traits examined may actually be component factors of others. Given this fact, it might be expected that we would observe clustering, or co-localisation of QTLs for various "related" traits on the genetic maps of the parents.

Instances of both the conservation of QTL effects across environments, and the co-localisation of QTLs for related traits are apparent in the current study. An excellent example of both phenomena is shown in Figure 2.18. In this case, a cluster of QTLs was detected on the same linkage group S-14(G) in both environments, with QTLs for leaf width and internode length in the field correlating to the position of QTLs for leaf width, leaf length and stolon length in the glasshouse. This might imply the presence of an environmentally stable genetic element influencing all of the associated traits. Another interesting phenomenon that was observed was the presence of a QTL cluster for a similar group of traits at a similar position on the other homoeologue group (S-15(G)) in this parent (Figure 2.18).

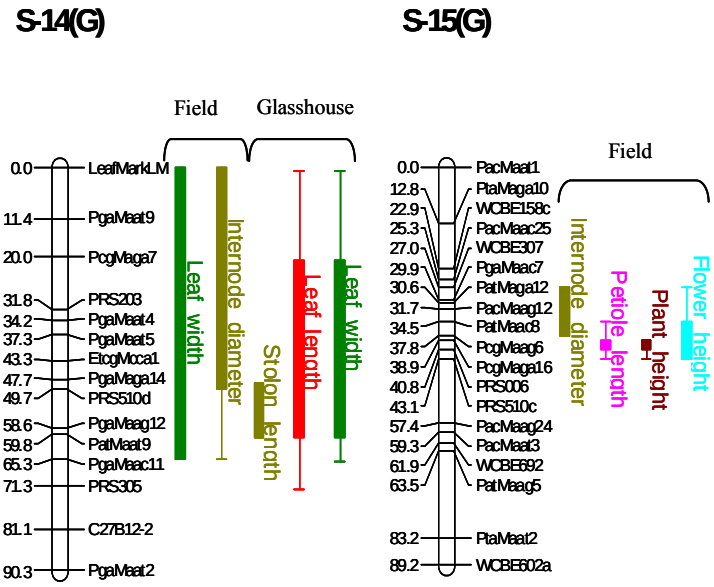


Figure 2.18. Linkage groups S-14(G) and S-15(G) showing similar QTL for leaf width both in the glasshouse and field experiment.

One interesting aspect of genome evolution in allotetraploidy is the possibility that homoeologous genes (on the two different "genomes") may either have retained the same function, or have become functionally divergent in nature because of the existence of independently segregating, redundant representatives of the same gene. In the former case (non-divergence of function), it might be expected that it would be possible for QTL for one trait to be observed on both homoeologues in one or both parents of a population. This may explain the co-incidence of QTL for similar and correlated traits on S-14(G) and S-15(G) (Figure 2.18). A similar phenomenon is observable for flowering time (Figure 2.19) where QTLs for this trait were observed

in a similar position on three of the four homoeologues of linkage group E. In this case, of course, the effect is conserved across both groups and parents, possibly indicating a locus that is consistently significant in the control of flowering time in white clover. In fact, on closer inspection of the QTL mapping data, a QTL for flowering time that falls just under the LOD cut-off threshold of 2.5 is present on R-9(E), supporting the possibility that this is an example of a gene/locus that has experienced functional conservation across homoeologous linkage groups.

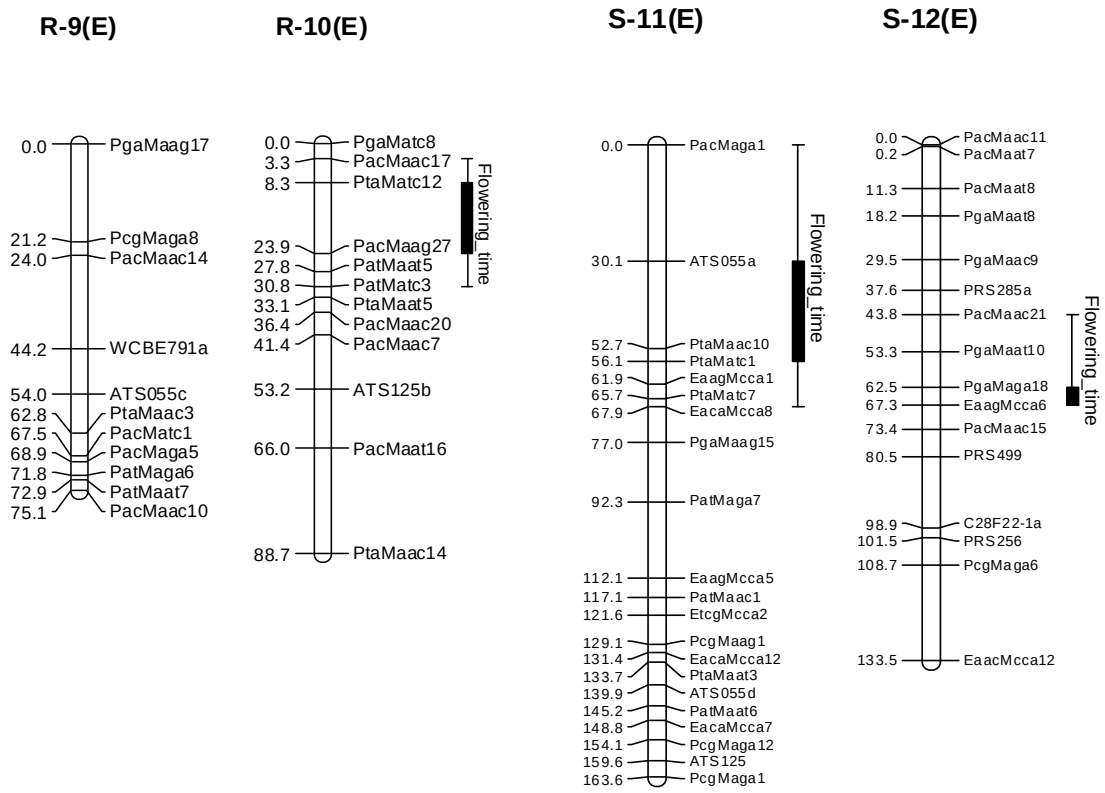


Figure 2.19. QTL for flowering time observed in a similar position on three of the four homoeologues of linkage group E for the field experiment.

A broad homology between the chromosomes of white clover and those of *Medicago truncatula* has been recently established (George *et al.*, 2006) and the homoeologous pair E in white clover is thought to correspond largely to *Medicago* chromosome 1. This chromosome harbours several genes involved in flowering pathways, including a gene of the blue light-absorbing phytochrome family (*CRY2b*) and the photoperiodic responsive flowering gene *GI* (Hecht *et al.*, 2005). Although it would be premature to make any definitive statement, it is interesting to speculate upon the possibility that such QTL for flowering time in clover might represent allelic variation in the clover homologues of genes such as these. A more detailed picture of the syntenic relationship between the two species might actually allow the

information in *Medicago* to be used to identify candidate genes underlying flowering time QTL in white clover. This concept is not restricted to flowering time – there are many mapped mutants and genes of known function in *Medicago*, which could act as a starting point for candidate gene identification for QTLs for various traits in clover given a sufficiently robust comparative genetics framework for the species.

One way in which the consistent involvement of a genetic interval in a particular trait can be established is to investigate the inheritance of the trait in multiple genetic backgrounds. Other mapping studies in white clover have been focused on QTL for seed production (Barrett *et al.*, 2005), for root-knot nematode resistance (Barrett *et al*. 2005a), and for morphogenetic and reproductive development traits (Cogan *et al.*, 2006). The latter, which used the genetic map from Jones *et al*. (2003), detected QTLs for a range of vegetative morphogenesis and reproductive morphogenesis and development traits in both individual and multi-environment combined analyses. Unfortunately, it is difficult to make comparisons between our results and the analysis carried out by Cogan *et al*. (2006), due to the lack of common markers between the two maps. The study by Barrett *et al*. (2005b) was focused on the genetic control of seed yield and its components in the F_1 mapping population used to construct the microsatellite map developed by Barrett *et al*. (2004). In this case the comparison of QTL location is more feasible as both the map from Barrett *et al*. (2004) and the map presented here comprise common markers. However, in this case, comparison can only be made between the relative positions of QTLs for different traits. For example, in the parental map R3R4, QTLs for two traits (petiole length and leaf width) analysed in the glasshouse experiment were found in a similar location of linkage group C as QTLs for two traits (seed yield and inflorescence density) detected by Barrett *et al*. (2005b). Similarly, in the parental map S1S4, QTLs for three traits (stolon length, leaf length and width) analysed in the glasshouse were detected in the same linkage of linkage group G as QTL for thousand-seed weight (Barrett *et al.*, 2005). The other study by Barrett *et al*. (2005b) used bulk-segregant analysis (Michelmore *et al*. 1991) to determine the map the QTL for root-knot nematode resistance. However the SSRs underlying this QTL were not successfully incorporated in the map presented in this study and no correlation could then be achieved.

Many of the most important agronomic traits in white clover are quantitative in nature. A greater understanding of all of these traits would doubtless benefit from a concerted effort to analyse the same traits, using the same markers, across multiple environments, in multiple populations. Already, several well characterised populations/maps exist, including the F_2 (I.4R x I.5J) population (Jones *et al.*, 2003), the F_1 (Sustain 65625/2 x NRS 364/7) population (Barrett *et al.*, 2004), the F_1 (TKPR x TNPR) population (Jones, 2005) and the F_1 (R3R4 x S1S4) population from this study. Key to the success of this strategy will be community-access to a large set of easily deployable, co-dominant genetic markers, which are preferably both homologue and homoeologue specific. It is hoped that collaborative initiatives such as the recently established International *Trifolium* Network (ITN) (http://www.trifoliumnetwork.org) will help the move towards this goal.

In addition to consolidating genetic studies in clover, it would be sensible to develop platforms to exploit the wealth of existing information in plants such as the recently sequenced model legume *Medicago truncatula*. One of the prerequisites for the efficient exploitation of *Medicago* genome sequence data for clover (and related functional genomics resources) is a good understanding of the level of the similarity between the genomes of the species on both macro- and microsyntenic scales. In order to contribute to the second of these goals, the remaining chapters in this thesis describe the construction of a BAC library in one of the parents (R3R4) of the mapping population used in this chapter, and a preliminary assessment of fine scale conservation of gene order and content between the two species using the BAC library and the genetic map described here.

3.0 Bacterial Artificial Chromosome (BAC) library construction and preliminary comparative sequence analysis with *Medicago truncatula*

3.1 Introduction

Over the past twenty years, much genome-based research has focused on model organisms in order to define the genetic architecture underlying key processes in a wide variety of organisms. For plants, *Arabidopsis thaliana* was chosen as the model for the dicotyledons due to its small genome size (125 Mb) (Meinke *et al.*, 1998a) and rice (*Oryza sativa*) was the first cereal to be sequenced (Nagamura *et al.*, 2003; Zhao *et al.*, 2004). However, in recent years investigators have sought a legume species that could serve as a functional genomic model for certain developmental systems that cannot be studied in *Arabidopsis* (Cook *et al.*, 1997a). Two model legumes, *Lotus japonicus* and *Medicago truncatula* have been adopted to understand the genetic architecture of legume species and to facilitate isolation and characterisation of genes responsible for legume-specific phenomena, including plant-microbe interactions such as symbiotic nitrogen fixation (Nakamura *et al.*, 2002).

An important aspect of transferring information from model to agronomically important species is knowledge of the extent of synteny between the two. Within the legumes, the extent of synteny has already been established, to varying degrees between *M. truncatula*, *M. sativa* (alfalfa), *L. japonicus* (Choi *et al.*, 2004b), *Glycine max* (Yan *et al.*, 2004), and *Pisum sativa* (pea). The extent of similarity between clover and these species remains largely unknown, although preliminary data available at the time of writing suggests that it is quite high. For example, George *et al*. (2006) have discovered a number of microsatellite-containing ESTs from white clover that have significant matches with the current *M. truncatula* and *L. japonicus* genome sequences, allowing the development of a comparative map that suggests high levels of macrosynteny between clover and *M. truncatula*.

As previously mentioned, white clover is a member of the tribe Trifolieae, which contains, amongst others, the closely related genera *Trifolium*, *Medicago* and *Melilotus.* Although white clover is a self incompatible allotetraploid, and *M. truncatula* is a self fertile diploid, both have the same basic number of chromosomes (x=8), and it is possible that gene order and organisation is largely conserved between the two species. Thus, the genome of *M. truncatula*, which is currently

being sequenced (Young *et al.*, 2005), may be useful as a tool to identify and isolate orthologous genes in the genome of white clover.

The main aim of this portion of the study was to construct a BAC library of white clover to use as a resource to assess the levels of fine-scale conservation of gene order (micro-synteny) between this species, *M. truncatula* and other model genomes. Prior to this, a pilot experiment was carried out by testing a set of *M. truncatula* PCR-based markers (originally developed by Choi *et al*. 2004a) in white clover to estimate the level of cross-amplification and to investigate the sequence-based similarity between the two species. In addition to the construction and basic characterisation of the library, we have sequenced both ends of over 700 BACs randomly chosen from the library. We present an analysis of these sequences with specific emphasis on the potential of the library for our proposed strategy to 'tile' the genome of white clover on to the genome sequence of *M. truncatula*. Our results suggest that, despite evidence for considerable rearrangements between the genomes of white clover and *Medicago truncatula*, large-scale BAC-end sequencing of the former has the potential to allow the anchoring of a significant portion of the genome of white clover on to that of the latter, significantly improving potential for gene discovery in white clover.

3.2 Material and Methods

3.2.1 Cross specific amplification of *Medicago truncatula* PCR-based markers in white clover

Choi *et al*. (2004a) constructed a core genetic map of *M. truncatula* enriched with gene-based genetic markers and they focused on three distinct classes of sequences:

(1) ESTs (expressed sequence tags) with high homology to genes known in Arabidopsis and/or other legume species. In cases in which introns could be predicted by aligning an *M. truncatula* EST with a corresponding genomic sequence of *Arabidopsis*, primer pairs were designed to anneal to exon sequences and to amplify across intron regions. In cases in which an *M. truncatula* EST possessed similarity to sequences identified in other legumes (on the basis of blastn), sequence alignments were used to design oligonucleotide primers that would amplify DNA fragments from each of the corresponding legume genomes.

(2) *M. truncatula* BAC clones with high homology to mapped soybean RFLP probes.

(3) Genes of predicted function, such as nucleotide binding site-leucine-rich repeat superfamily of resistance gene analogues and with a possible role in plant-microbe interactions, including symbiotic nitrogen fixation and pathogenic associations.

For the course of this study, 95 of those PCR-based markers were analysed on the white clover mapping parents (S1S4 and R3R4) (Appendix B). PCR was performed on each of the primers on the two white clover parents as well as *M. truncatula* as positive control. Annealing temperature and magnesium chloride ($MgCl_2$) concentrations were optimized for each of the primers and a standard protocol has been developed (Table 3.1 and 3.2).

Table 3.1. Standard PCR components used for *M. truncatula* PCR-based markers.

Component	Volume/single reaction	Final concentration
DNA sample (25ng/µl)	2	2.5 ng/20 µl
10X Buffer	2	1X
dNTPs (2 mM)	2	200 µM
$MgCl_2$ (10 mM)	2	2.5 mM
Upstream primer (10 µM)	0.2	0.1 µM
Downstream primer (10 µM)	0.2	0.1 µM
Taq polymerase (5 u/µl)	0.2	0.05 u/20 µl
dH_2O	11.4	To a final volume of 20 µl

Table 3.2. Amplification condition for standard PCR used for *M. truncatula* PCR-based markers.

Initial Denaturation	94°C x 3 min	1 cycle
Amplification:	94°C x 0.45min	
	55°C x 0.45min	35 cycles
	72°C x 1.30min	
Final extension	72°C x 10min	1 cycle
Hold temperature	4°C	

Each of the *M. truncatula* PCR-based markers that resulted in single copy amplicons in white clover was also cloned according to the TA Cloning® kit (Invitrogen). The PCR products were ligated with pCR® 2.1 cloning vector (Figure 3.1) and incubated at 14°C overnight. The ligation reaction was then transformed with One Shot® Competent TOP10 Cells (Invitrogen) onto plates containing Luria Bertani (LB) agar with kanamycin (50 mg/ml) and X-Gal (40 mg/ml). The plates were incubated at 37°C overnight. A colony PCR was performed on four of the colonies on each plate to check for the presence of the PCR product as insert. The positive colonies were stabbed onto LB^{+K} plates and sent for direct sequencing to Agowa (Germany).

The resulting sequences were compared to the original *Medicago truncatula* sequences using the BLAST 2 Sequences tool (http://www.ncbi.nlm.nih.gov /blast/bl2seq/wblast2.cgi) with an e-value cutoff of e-8 and a percent identity of 40%.

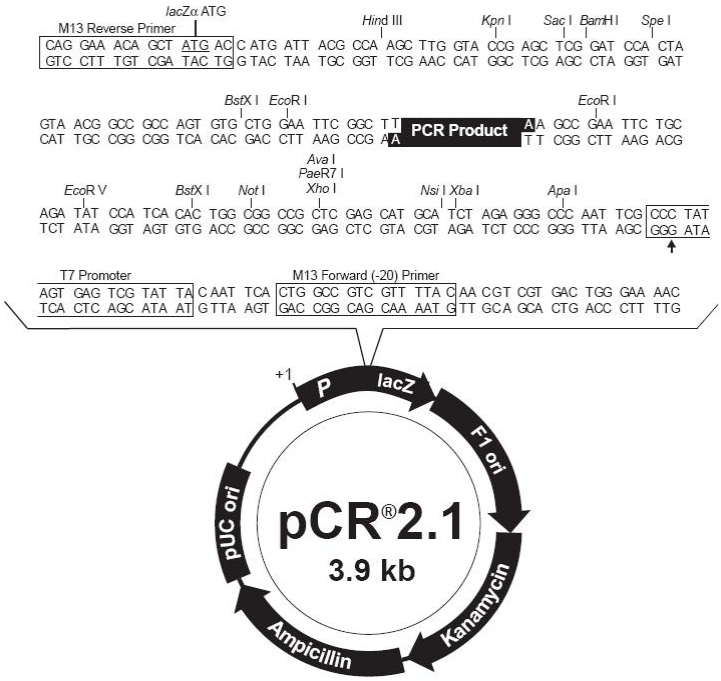


Figure 3.1. Map of the cloning vector pCR®2.1 (Invitrogen) used to clone the single copy amplicons.

3.2.2 Preparation of insert DNA

3.2.2.1 Isolation of high molecular weight nuclear DNA

- **Isolation of nuclei**

Cuttings of white clover plants (R3R4) were transferred to 10 cm pots of sterile soil and grown in the glasshouse until 20-30 cm tall. For each round of nuclei isolation, two or three selected plants (not more than 4 weeks old) were kept in the dark for at least 48 hours. 10 g of leaf material was harvested, and roughly chopped up with a clean razor/scalpel blade. The chopped leaves were swirled for 30 seconds in diethyl ether and washed 3 times in 100 ml of dH_2O. 100 ml of Honda buffer (34 mM Tris-HCl pH 8.0, 7 mM $MgCl_2$, 628 mM Sucrose, 0.69 mM spermidine, 2.7 mM spermine, 3.41% v/v Triton X-100, 6.83% w/v Dextran T40, 3.41% w/v Ficoll PM400) was added and the mixture blended with 2 x 5 seconds pulses. The resulting suspension was gravity-filtered through 4-8 layers of Miracloth into a sterile container. The filtrate was transferred into 2 x 50 ml centrifuge tubes and centrifuged at 1800 g, 4^{o}C, for 15 min and then left to stop with brake at lowest. The supernatant was discarded and the pellet was resuspended in Yamaha buffer (Honda buffer without Dextran T40, Ficoll PM400 and Triton X-100) using a small sterile paintbrush. Both samples were bulked into one tube, the volume was made up to 50

ml with Yamaha buffer and the previous spin was repeated. If the supernatant is slightly green, the pellet was resuspended again and the previous step repeated. The clear pellet was then resuspended in a final volume of 400 µl Yamaha buffer.

- **Preparation of plugs**

10 ml of 1.2% LMP (low melting point) agarose was made up in Yamaha buffer. Once the mixture was dissolved, it was placed in a 50^{o}C water bath. An equal volume (400 µl) of the LMP agarose solution with pre-warmed nuclei (10 min at 50^{o}C), and the mixture was placed into the wells of a pre-chilled plug mould. The plugs were allowed to solidify in the refrigerator for 30 minutes. The plugs were then pushed out of plug into 50 ml of Lysis Buffer (1% w/v sodium lauryl sarcosine, 0.1 mg/ml proteinase K (added just before use) dissolved in 0.5 M EDTA pH 9.2). The plugs in Lysis Buffer were incubated at 50^{o}C for 48 hours.

3.2.2.2 DNA analysis

The quantification of the DNA in the plugs was carried out by placing 0.5, 0.25, 0.125 of one plug in a 1% agarose/0.5X TBE gel using the CHEF gel apparatus. PFGE Lambda Ladder was used as standard ladder and 10X uncut lambda DNA at various concentrations (10 µg, 5 µg, 2.5 µg) was used as sizing standard. The gel was run using the following parameters: 1 sec initial switch time, 40 sec final switch time, 18 hours run time, 6 Volts/cm, 120^{o} included angle and linear ramping. The gel was then stained and its UV fluorescent image was captured.

3.2.2.3 Test restriction digest

The agarose plugs containing high molecular weight DNA obtained in Section 2.7.2.1 were in 0.5 M EDTA. The plugs were incubated twice for an hour in 50 ml of 1X TE, 0.1 mM PMSF (Phenylmethylsulphonylfluoride) and were then washed 4 times for 30 minutes in 1X TE. Generally 4 X 100 µl plugs were chopped to a fine granular suspension on a clean microscope slide using a clean razor blade. The chopped plugs were placed in a 1.5 ml tube with a TE/Triton X-100 solution (1X TE, 0.1% Triton X-100). The agarose was resuspended by vortexing for 3 to 5 seconds and then subjected to a 3-5 seconds pulse on a benchtop microcentrifuge at full speed, yielding to a liquid phase on top of a concentrated and consistent suspension

of agarose. 15 µl of a restriction buffer mix (0.43 X *Hin*dIII Restriction Buffer, 14 mM Spermidine, 0.47 mg/ml BSA) was added to each tube containing 50 µl of the agarose suspension. The tubes were placed on ice for 30 minutes to equilibrate. A serial dilution (Table 3.4) of the restriction enzyme *Hin*d*III* was used to estimate the best conditions for partial digest of a specific batch of DNA.

Table 3.4. This table shows the *Hin*dIII serial dilution used for the test restriction digest.

Dilution	Volume of HindIII	Volume of 1X Buffer (µl)	Final concentration in 50 µl aliquots
1	2 µl of HindIII (20 units)	23	4
2	2 µl of dilution 1	18	2
3	2 µl of dilution 2	18	1
4	2 µl of dilution 3	18	0.5
5	2 µl of dilution 4	18	0.25

One enzyme dilution was added to each tube (except for the undigested control) and placed on ice to equilibrate. The tubes were then incubated at 37^{o}C for 30 minutes and the reaction was stopped by adding 0.135 M EDTA to each tube. After 30 minutes on ice, each reaction was loaded on a 1% agarose/ 1X TBE gel using the CHEF gel apparatus. PFGE Lambda Ladder was used as standard ladder and 10X uncut lambda DNA at various concentrations (10 µg, 5 µg, 2.5 µg) was used as sizing standard. The gel was run using the following parameters: 1 sec initial switch time, 20 second final switch time, 18 hours run time, 6 Volts/cm, 120^{o} included angle and linear ramping. The gel was then stained and its UV fluorescent image was captured.

3.2.2.4 Mass digestion and first size selection

In scaling up the partial digest, the above experiment (Section 3.1.2.3) was repeated with the optimal enzyme concentration, which yielded the brightest smear of digested DNA in the desired size range (100 to 300 Kb).

On a 1% agarose gel/ 0.25X TBE using the Biorad DR®II CHEF Gel Apparatus, a slot well was prepared so that the 4-6 wells in the centre of the comb were taped together to produce a single well capable of containing 400 µl. The macerated, digested plug pieces were transferred into the slot well. The PFGE Lambda Ladder was used as ladder and all the wells were sealed with melted agarose. The gel was

run using the following parameters: 6 volts/cm, 120^{o} included angle, 1 sec initial switch time, 20 sec final switch time, linear ramping, 16 hours run time, 12^{o}C buffer temperature. On completion of the run, the portions containing the markers and approximately 2 mm of the large well were excised and stained. After staining, these gel portions were photographed on a UV transluminator, placing a UV-fluorescent ruler beside the gel slice (0 cm = wells of the gel). The position of the segment of the gel representing the size range from 150 to 250 Kb was measured relative to the wells.

3.2.2.5 Second size selection

Early experiments based on a single size selection step resulted in BAC libraries with an average insert size below expected. This was considered to be due to the phenomenon of "trapping", where smaller DNA molecules were unable to migrate faster than larger ones due to the high concentration of DNA being electrophoresed. In order to ameliorate this effect, a second size selection was performed.

The region of unstained CHEF gel containing the partial digested DNA in the desired size range was located and cut out of the gel using the measurements taken above. The gel slice was then placed on a gel-casting tray in reverse orientation and embedded in a new 1% agarose gel. This gel was run using identical conditions as previously described and the position of the size selected DNA again determined by staining portions of the gel containing only the edges of the migrated DNA, and measuring its position relative to the bottom of the gel. The effect of reversing the orientation of the gel slice was that the second size selected DNA was compressed into a smaller area than the original size selection, allowing recovery of large DNA at a concentration which allowed efficient ligation into the vector.

3.2.2.6 Isolation of size-selected DNA from agarose

Excision of the gel slice holding the partially digested DNA was performed using the punch portions of Quick-Pik® electroelution capsules from Stratagene®. The assembled capsules were submerged in 1X TAE (40 mM Tris, 20 mM Acetic acid, 1 mM EDTA) in an electrophoresis tank with the cup facing the positive terminal and the punch facing the negative terminal. The run was performed with a voltage of around 5 Volts/cm to electroelute the DNA. After 3 hours the current was reversed

for 30 seconds. Approximately 50 µl of electroeluted DNA was collected from each capsule using a cut-off pipette tip. The DNA samples were placed on ice and a 1% agarose submarine mini-gel was prepared in 1X TAE. A set of 1X uncut lambda DNA was prepared at different concentrations (25 ng, 50 ng, 100 ng and 200 ng). The sizing standards and 5 µl of the DNA samples containing blue juice were loaded and the gel was run at 100V for 15-20 minutes. Based on comparison of the relative fluorescence in the sample and standard lanes, an estimate of the concentration of each sample can be made. Multiplication of a sample's volume by its concentration gives an estimate of the total amount of insert DNA in that sample.

3.2.3 Library construction

3.2.3.1 Cloning vector, *pIndigoBAC-5*

pIndigoBAC-5™ (Figure 3.2, Epicentre®) is a 7506 bp bacterial artificial chromosome (BAC) cloning vector. It is derived from pIndigoBAC and pBeloBAC11 (Shizuya *et al.*, 1992). This vector has a enhanced color intensity production in X-gal-based "blue-white" screening and comprises a chloramphenicol resistance site, E. coli F factor partitioning and copy number regulation system, bacteriophage lambda *cos* site for lambda packaging or lambda-terminase cleavage, bacteriophage P1 loxP site for Cre-recombinase cleavage, bacteriophage T7 RNA polymerase promoter flanking the cloning site, BAC-end sequencing primer binding sites.

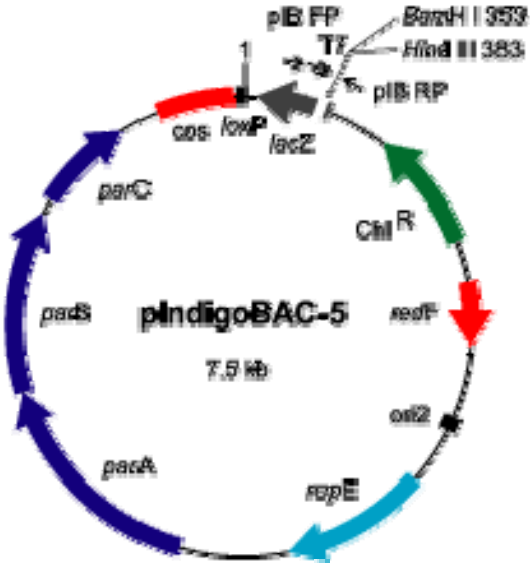


Figure 3.2. pIndigoBAC-5 cloning vector used in the white clover BAC library.

3.2.3.2 Ligation

The partially digested high molecular weight DNA was ligated into the BAC vector pIndigoBAC-5 (*Hin*dIII- Cloning ready, Epicentre®). The ligation reactions were set up as follows: 1.25 μg vector DNA, 1X T4 ligase buffer (New England Biolabs), 6 units T4 ligase (Promega), 100 ng insert DNA. The reaction tubes were incubated at 16^{o}C overnight and then they were placed in a 65^{o}C water bath for 20-30 minutes to heat-kill the enzyme.

If immediate transformation and colony picking was possible, ligations were stored at 4^{o}C. However storage of ligations for period of greater than 2 weeks resulted in vastly reduced transformation efficiencies. Therefore, for long-term storage, ligations were divided into 5 μl aliquots, flash frozen in liquid N_2 and placed immediately at –80^{o}C, where they would be maintained for a period of 2 months without any significant drop in transformation efficiency.

3.2.3.3 Transformation

If flash frozen, ligations were thawed on ice before transformation. Electroporation cuvettes (1 mm gap, 100 μl volume, Eppendorf) were chilled on ice for 5 to 10 minutes. ElectroMax^{TM} DH10B^{TM} cells (Invitrogen^{TM}) were thawed on ice and 20 μl of thawed cells were added to 1 μl of chilled ligation reaction and mixed gently using a cut-off pipette tip. The tubes were left on ice for 2-3 minutes; the DNA/cells mix pipetted into a chilled cuvette and tapped down to the bottom. After a successful electrical pulse using the Electroporator 2510 (Eppendorf) on a setting of 1800 Volts, 1 ml of SOC medium (20 g/L peptone, 5 g/L yeast extract, 0.5 g/L NaCl, 2.5 mM KCl, and just before use 10 mM $MgCl_2$, 20 mM glucose) was added to the cuvette and mixed gently. The SOC/cells mix was transferred into a sterile 50 ml tube. The tube was incubated at 37^{o}C for 1 hour. 100 μl of the solution was spread onto a room temperature X/I/C plate (20g/L LB, 1.2% w/v technical agar, 90 mg/ml X-Gal, 90 mg/ml IPTG, 12.5 μg/ml chloramphenicol). The plates were incubated at 37^{o}C overnight and then placed at 4^{o}C for 1-3 hours to allow blue/white color selection to develop sufficiently.

3.2.3.4 Miniprep and *Not*I digests

Individual recombinants clones (white colonies) were transferred into a 5 ml LB/Chloramphenicol (12.5 μg/ml) using sterile cocktail sticks. When all the white colonies have been transferred, the resulting cultures were incubated overnight in a shaker at 37^{o}C. Subsequently 2 ml of cell culture was centrifuged at 14000 rpm for 1 minute in a benchtop centrifuge. The bacterial pellets were resuspended in 200 μl of MP-1 (50 mM glucose, 25 mM Tris HCl, pH 8.0, 10 mM EDTA) and incubated on ice for 5 minutes. 400 μl of MP-2 (0.2 M NaOH, 1% w/v SDS) was added and the tube gently mixed. The mixture should turn translucent due to lysis of the bacteria. The tubes were placed on ice for 5 minutes and 300 μl of MP-3 (3 M potassium acetate, 28.5% v/v acetic acid) was added and a white precipitate should appear, which contains genomic DNA and various proteins. The tubes were placed on ice for 7 minutes and then centrifuged at 13,000 rpm for 25 minutes. The supernatant of each tube was transferred into clean tubes. To precipitate plasmid (BAC) DNA 600 μl of ice-cold isopropanol was added and the tubes placed at –80^{o}C for 20 minutes. The tubes were then spun at 13,000 rpm for 30 minutes, the supernatants were carefully poured off and 1 ml of ice-cold 70% ethanol was added to each tube. The tubes were centrifuged at 13,000 rpm for 10 minutes. Once the supernatant has been discarded, the translucent pellet (BAC DNA) was allowed to dry and then 35 μl of 1X TE was added and the tubes were placed overnight at 4^{o}C.

To release the insert, BAC DNA was digested with *Not*I (New England Biolabs). The 20 μl reactions contained 5 units of *Not*I, 1X Buffer, 4 mM spermidine, and 5 μl of DNA from the above miniprep. The reactions tubes were centrifuged at 10,000 rpm for 30 seconds, placed into a rack, wrapped with aluminium foil and placed at 37^{o}C for 1 hour. The digests were run onto a CHEF Gel Apparatus (BioRad). The ladder DNA (PFGE Lambda Ladder, New Englands Biolabs) is embedded in agarose loaded in a 1.0 ml syringe. The gel was run using the following parameters: 6.0 V/cm, 120^{o} included angle, 16 hours run time, 5 sec initial switch time, 15 sec final switch time, linear ramping. The gel was then allowed to stain in 10 mg/ml ethidium bromide for 45 min, and destained for another 45 min. The gel image was photographed under UV light.

3.2.3.5 Establishing and storing the BAC library

Once the library has been plated, clones are picked and placed into 60 µl freezing media [2.5% w/v granulated LB agar (EM Science), 13 mM KH_2PO_4, 36 mM K_2HPO_4, 1.7 mM sodium citrate, 6.8 mM $(NH_4)_2SO_4$, 4.4% v/v glycerol, 12.5 µg/ml Chloramphenicol] in microtiter plates to create an ordered BAC library. Each suitable white colony is placed into a single well of a microtiter plate (384-well plate) using a toothpick. The plates were incubated at 37^{o}C for 14-16 hours. This constitutes the master copy of the BAC library.

For every 384-well plate containing the master copy of the library, one 384-well plate was filled with 75 µl freezing media. Replication was carried out manually using 384-pin replicators (Genetix). For storage, the 384-well plates were arranged in stacks of nine plates and the master copy of the library was stored into a box into a –80^{o}C. The same was carried out for the two other copies.

3.2.4 Characterisation of the BAC library

3.2.4.1 Average insert size

To determine the average insert size of the library, DNA from 100 randomly selected BAC clones was isolated and digested with *Not*I as described in Section 3.2.2.4. The average insert size was calculated based on the PFGE ladder and a distribution graph was plotted.

3.2.4.2 PCR screening of the library

Plate pools were constructed in order to test the coverage of the library by determining the number of times a group of single copy sequences were present in the library. This was achieved by replicating each of the plates of bacterial clones onto an individual LB^{+CM} agar plate. The replication was performed using 384 pin manual replicators (Genetix). Agar plates were incubated at 37^{o}C overnight. Subsequently 4 ml of LB broth were added to each plate into which the colonies were scraped off using a sterile spreader and removed into 2 x 2 ml tubes. One set of tubes was centrifuged at 13,000 rpm for 10 minutes and the supernatant was discarded. The tubes were then placed at –80^{o}C for storage. The other set of tubes was subjected to an alkaline lysis method to isolate the BAC DNA. The tubes were

centrifuged at 13,000 rpm for 10 minutes and the supernatant was discarded. The resulting pellet was submitted to BAC DNA miniprep as described in Section 3.2.3.4, with the exception that the resulting BAC DNA pellet was resuspended into 150 µl of 1X TE buffer. The concentration of the BAC DNA samples was quantified by electrophoresis against a series of lambda DNA standards on a 1% agarose gel. The concentration was adjusted to 25 ng/µl for further analysis.

The plate pools were then screened by PCR amplification with four white clover SSRs (ATS123, PRS256, ATS176, ATS055) (Chapter 2, Section 2.2.4.1) (Barrett *et al.*, 2004) and five PCR-based markers from *Medicago truncatula* (DK501R, DNABP, RNAH, 5J9L, BE187590) (see Section 3.2.1) (Choi *et al.*, 2004a).

3.2.4.3 Estimation of chloroplast contamination of the library

Screening for chloroplast contamination was performed by PCR using consensus chloroplast (Cc) primer pairs (Chung & Staub, 2003). Three primer pairs CcSSR6, CcSSR15, CcSSR22 were chosen to screen a set of randomly selected BAC clones. These CcSSR are contained in genes based on the tobacco chloroplast genome (RpoB, Rpl20-ClpP, TrnL-16SrRNA respectively) and are evenly distributed across the chloroplast genome, with intervals of approximately 50 to 60 Kb between each primers (position according to the tobacco chloroplast genome, accession number = CHNTXX). The BAC clones were incubated in a 384-well plate containing 10 µl LB with 12.5 µg/ml Cm and grown at 37^{o}C overnight, and 1 µl of cultures was added into a 10 µl multiplex PCR reaction (Table 3.5 and 3.6). After PCR the products were analysed in a 2% agarose gel and the chloroplast contamination determined by the frequency of occurrence of the expected amplicons for the primer sets.

Table 3.5. Standard PCR components used to estimate the chloroplast contamination of the BAC library.

Component	Volume/single reaction	Final concentration
BAC clone culture	1	1 µl culture / 10 µl
10X Buffer	2	1X
dNTPs (2 mM)	1.5	300 µM
$MgCl_2$ (10 mM)	2	200 µM
Upstream primer (10 µM)	0.3	0.03 µM
Downstream primer (10 µM)	0.3	0.03 µM
Taq polymerase (5 u/µl)	0.4	0.08 u/10 µl
dH_2O	4.5	To a final volume of 10 µl

Table 3.6. Amplification conditions used for the PCR reaction used to estimate the chloroplast contamination of the BAC library.

Initial Denaturation	94°C x 5 min	1 cycle
Amplification	94°C x 1 min 50°C x 1 min 72°C x 1 min	35 cycles
Final extension	72°C x 6 min	1 cycle
Hold temperature	4°C	

3.2.5 BAC-end sequencing analysis

BAC-end sequencing involves obtaining a single sequence read from both ends of the DNA insert. BAC-end sequence data is collected from a large number of randomly selected clones in a given library, and is used to infer the genomic potential and diversity of the large insert library (and thereby, the original sample).

In this study, the BAC-end sequence analysis was performed with collaboration with Prof N.D. Young and Dr. S. Cannon at the University of Minnesota (St Paul, MN) and Dr C.D. Town and F. Cheung at the Institute for Genomic Research (TIGR, Rockville, MD).

DNA template was prepared in 384-well format by a standard alkaline lysis method. End sequencing was performed using Applied Biosystems (ABI) Big Dye terminator chemistry and analysed on ABI 3730xl machines. Base calling uses TraceTuner and sequences are trimmed for vector and low quality sequences using LUCY (Chou & Holmes, 2001). Sequences were searched for protein-coding regions in the TIGR Plant Gene Indices (http://www.tigr.org/plantProjects.shtml) using BLASTn with a cut-off value of 1e-20 (Quackenbush *et al.*, 2000) and using BLASTx (cut-off value 1e-10) in the TIGR non-identical amino acids database that contains non-identical protein data from GenBank, RefSeq, Uniprot, CMR, PDB, and PRF. The BAC-end sequences were also compared with repetitive DNA in the TIGR Transposon database using BLASTx with a cut-off value of 1e-10. The BAC-end sequences were compared with the *Medicago* genome sequence contig (http://www.tigr.org/tdb/e2k1/mta1/) using BLASTn with a cut-off value of 1e-10. To identify clover BACs that were likely collinear (i.e. showed microsynteny) with the *Medicago truncatula* genome, on the basis of 167,690,648 bp of *Medicago* Genomic sequence, the searches against the *Medicago* genomic sequence were parsed first to remove

transposon matches, and then to identify BACs for which both ends had a significant match to a stretch of *Medicago* sequence and where the two regions on the *Medicago* genome were between 0 and 200 Kb apart.

3.2.6 Development of microsatellites from the BAC-end sequences

The BAC-end sequences containing simple sequence repeats were identified using the computer programme MISA (Appendix B) (Thiel *et al.*, 2003). Primers for BESs containing SSRs were designed using the Primer 3 software (Appendix C). Each primer pair was tested for amplification by PCR on the genotype used for the BAC library construction (R3R4) and another genotype (S1S4). Both these constitute a mapping family with 94 progeny. The primer pairs were then tested for polymorphism using radioactive labelling in the two mapping parents and 4 F_1 progeny (Table 3.9 and 3.10). The PCR products were analysed on a 5% polyacrylamide gel and the gels were dried on Whatmann 3MM paper and exposed to storage phosphor screens for 1 to 3 days at room temperature.

Table 3.9. Standard PCR components used for BES microsatellites.

Component	Volume/single reaction	Final concentration
DNA sample (12.5 ng/µl)	1	1.25 ng/10 µl
10X Buffer	1	1X
dNTPs (2 mM)	1	200 µM
[^{33}P]-ATP labelled primer	0.25	0.25 µM
Non labelled primer (10 µM)	0.1	0.1 µM
Taq polymerase (5 u/µl)	0.05	0.025 u/10 µl
dH_2O	6.6	To a final volume of 10 µl

Table 3.10. Amplification condition for standard PCR used for BES microsatellites.

Initial Denaturation	94°C x 3 min	1 cycle
Amplification:	94°C x 0.45min	
	50°C x 0.45min	34 cycles
	72°C x 0.45min	
Final extension	72°C x 2min	1 cycle
Hold temperature	4°C	

The polymorphic SSR primer pairs were then amplified in the mapping population using fluorescent labelling (See Chapter 2, Section 2.2.4.2), run on the genetic analyser ABI3100® (Applied Biosystems) and analysed on the ABI Prism® GeneMapper™ Software Version 3.0. The segregating SSR markers were then mapped onto the genetic linkage map described in Chapter 2.

3.3 Results

3.3.1 Cross specific amplification of *M. truncatula* PCR-based markers in white clover

A set of 95 *M. truncatula* PCR-based markers developed by Choi *et al*. (2004a) was tested for amplification in white clover. The aim of this experiment was to obtain an initial estimate of the sequence-based similarity of white clover and *M. truncatula*, and to investigate a possible source of genetic markers for clover. 89 (94%) of the markers tested amplified in *M. truncatula* and 65 (73%) amplified in the white clover parents (R3R4, S1S4). The relatively high level of amplification from one species to the other is a first indicator to how similar the two species are in terms of sequence content, although it says little about levels of gene order conservation.

The EST-based markers were chosen to represent apparently low- or single-copy-number genes using the *Arabidopsis* genome as a reference for gene copy number. The markers derived from *M. truncatula* BAC clones with high homology to mapped soybean RFLP probes were selected to appear in majority as a single locus in *M. truncatula*. However of the 65 primers that amplified in white clover, only 22 amplified as a single copy (Figure 3.3); the remaining amplified in a more complex pattern (two or more amplicons). This low percentage (34%) of single copy amplification is not entirely surprising given that white clover is a supposed allotetraploid, a factor which could easily yield multiple products derived from the two separate homoeologous genomes. However, in some cases, the amplification patterns of markers in clover were quite complex, suggesting to a greater degree of complexity in white clover than could be accounted for by tetraploidy.

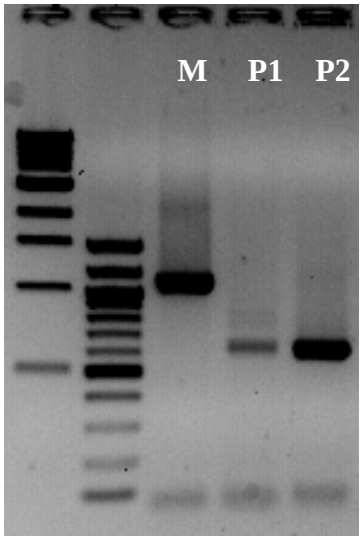


Figure 3.3. Amplification of *M. truncatula* (M) and the white clover parents (P1, P2) with RBBP primer. Lane 1 and 2 represents 1 Kb and 100 bp ladder respectively.

To obtain a better idea of the overall level of sequence similarity between white clover and *Medicago truncatula*, the 22 PCR-based markers that gave single amplicons in white clover were amplified in both mapping parents (R3R4 and S1S4) and were cloned into pCR® 2.1 cloning vector. Of these single amplicons, 17 (81%) were sequenced and their direct PCR sequences were compared to the corresponding *Medicago truncatula* sequences using BLAST-2-sequences software (Table 3.11). The table shows that 13 of the 17 white clover sequences (76.5%) had a high homology with the *M. truncatula* sequences and two of these were only similar to the mapping parent R3R4. The remainder four white clover sequences contained a viable insert but with no similarity to *M. truncatula* sequences. The *M. truncatula* markers originating from BAC-end sequences showed a high similarity with white clover with the exception of one marker (26G3L). The other two classes of markers derived from EST sequences showed a weaker similarity and quite often discontinuous when more than one part of the white clover sequence had similarity with *M. truncatula* (Figure 3.4), a situation which may be due to the divergent sequence content (and possibly, the position) of introns between *M. truncatula* and clover.

Table 3.11. Results of the BLAST 2 sequences analysis between the white clover mapping parents (R3R4 and S1S4) for the PCR-based markers from Choi *et al*. (2004).

***M. truncatula* markers**	**Marker type**	**R3R4**		**S1S4**	
		E-value	**% Identity**	**E-value**	**% Identity**
7H15L	BEST*	2 e-74	82	8 e-74	83
DK501R		1 e-97	88	1 e-94	88
DK298R		1 e-158	90	2 e-134	87
DK009R		0.0	93	0.0	94
40L12R		2 e-78	86	7 e-82	86
5J9L		2 e-137	92	1 e-128	91
26G3L		-	-	-	-
EST758	ESTe*	3 e-23 / 9 e-20**	95 / 96**	-	-
VBP1		1 e-109 / 2 e-08**	94 / 100**	2 e-78 / 2 e-08**	93 / 100**
BE187590		4 e-45 / 4 e-12 / 5 e-04**	78 / 86 / 96**	1 e-59 / 0.028**	77 / 93**
UNK7		-	-	-	-
DNABP	ESTi*	2 e-15	88	1 e-21	92
RNAH		3 e-73 / 1 e-06**	96 / 97**	3 e-73 / 1 e-06**	96 / 97**
NCAS		3 e-108 / 0.086**	88 / 100**	3 e-115 / 0.086**	89 / 100**
AW736703		-	-	-	-
AW257289		-	-	-	-
RBBP		4 e-55 / 2 e-13 / 2 e-06**	84 / 89 / 93**	-	-

* Marker types according to Choi *et al*. (2004): BEST = BAC-end sequence-tagged markers, ESTe = exon-derived markers, ESTi = exon-derived/ intron spanning markers.

**The slashes (/) represent the discontinuity in sequence similarity between white clover and *M. truncatula*.

Figure 3.4. Results of the Blast2Sequence analysis of the sequence of *M. truncatula* marker EST758 (2) and the white clover amplicon sequence (1). The picture shows the discontinuity in the sequence similarity.

3.3.2 BAC library construction

The white clover R3R4 BAC library was constructed from high molecular weight (HMW) DNA isolated from nuclei using the Honda method and embedded in agarose. Quantification of nuclear DNA isolated was determined on a CHEF gel and was estimated at megabase size DNA of sufficient quality and quantity for BAC library construction (Figure 3.5, A).

Nuclear DNA was released from agarose using QuickPik capsules before partial digestion with *Hin*dIII to ensure the exposure of all DNA to enzyme. Based on the analysis of ethidium bromide-stained pulsed-field gels, 0.5 units of *Hin*dIII produced the largest amount of DNA in the 50-200 Kb size range (Figure 3.5, B). After two size selections, the library was constructed using the *Hin*dIII site of the vector pIndigo BAC-5. Five separate ligations gave rise to the library consisting of 37,248 clones arrayed in 97 x 384 well microtiter plates. The master copy and two other copies of the library are stored at –80^{o}C in separate freezers.

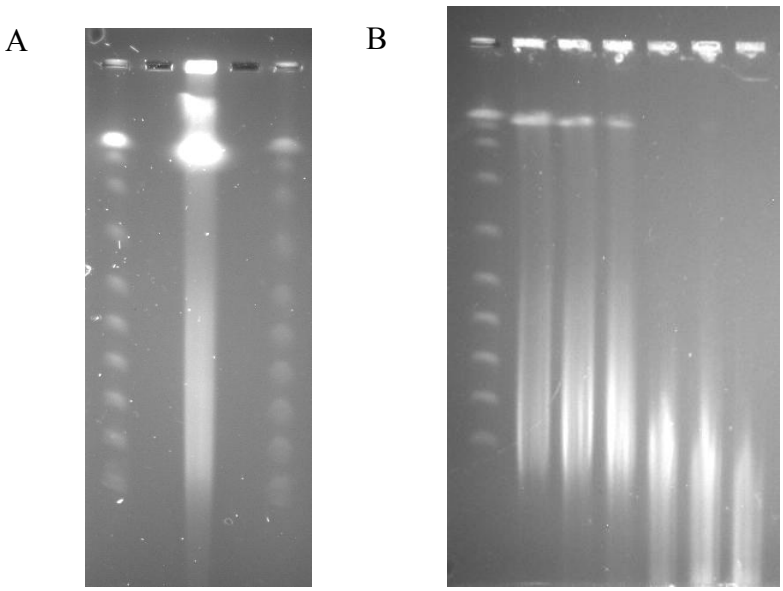


Figure 3.5. A: Pulse-field gel electrophoresis (PFGE) representing the HMW DNA isolation from nuclei. B: PFGE representing the test restriction digest with *Hin*dIII in decreasing concentrations (4, 2, 1, 0.5, 0.25, 0 units). PFGE Lambda Ladder was used as standard ladder.

3.3.3 BAC library characterisation

3.3.3.1 Average insert size

In order to determine the average insert size of the library, DNA from 86 randomly selected BAC clones was isolated and digested with *Not*I enzyme to release the DNA insert from the cloning vector and analysed by pulse-field gel electrophoresis (PFGE). The average insert size was determined by calculating the average between the sizes of all the 86 individual clones. Out of the 86 BAC clones tested, all contained white clover insert-DNA. The insert sizes ranged from 45 to 146 Kb with an average of 82 Kb (Figure 3.6). The size distribution of these clones (Figure 3.7) appears to represent a normal distribution, with the majority of clones in the 80 to 90 Kb size range.

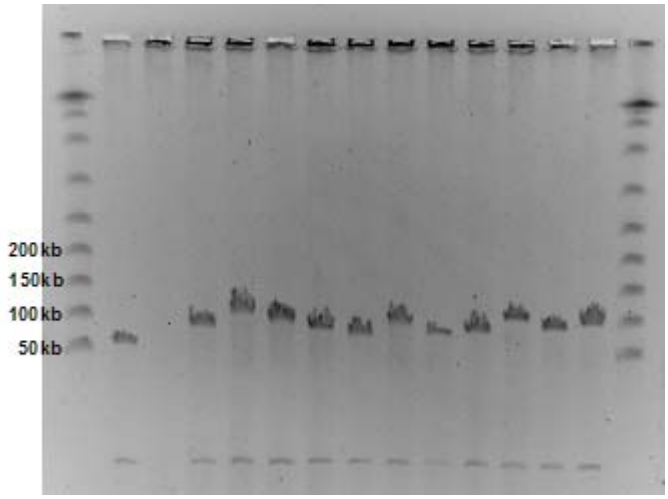


Figure 3.6. An analysis of white clover BAC clones by PFGE. Lane 1, 15 Lambda ladder, Lanes 2-14 Fragments of *Not*I-digested DNA isolated from randomly selected BAC clones.

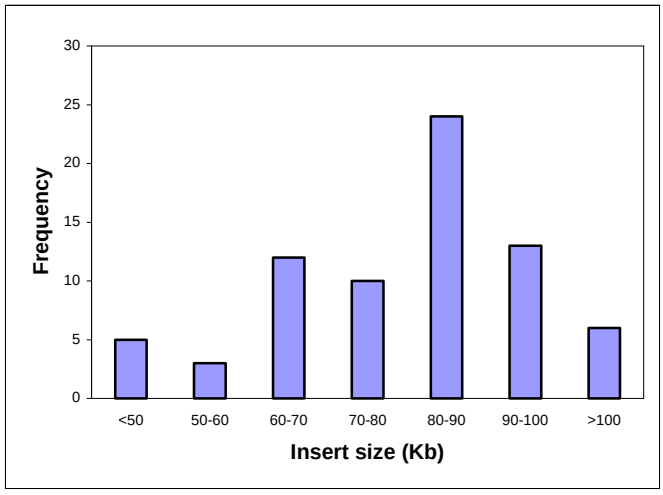


Figure 3.7. Distribution of insert sizes of randomly selected BAC clones. The insert size is plotted versus its frequency.

3.3.3.2 Screening of the library

A plate pooling strategy allowing the library to be screened by PCR was used to confirm the coverage of the library (Section 3.2.3.2.). The 97 plate pools were assayed by PCR using four white clover microsatellite primers previously mapped on the white clover map and five PCR-based markers from *Medicago truncatula* PCR-based primers from Choi *et al* (2004a) (Table 3.12, Figure 3.8). The occurrence of the appropriate PCR product for a pool after electrophoresis indicated that at least one of the 384 cultures in the corresponding library plate contained a BAC with the target sequence. Screening of these markers resulted in a range of 1 to 9 hits, with an average of 5.7 hits per marker (SD ±2.65) (Table 3.12). Given the allotetraploid nature of white clover, with each single copy marker having two potential homoeologues in the genome, and assuming that each hit in a BAC-pool represents only one positive BAC in that pool, an average of approximately 6 hits per marker is completely consistent with the estimated three fold genome coverage of the library.

Table 3.12. Number of hits in the plate pools of the R3R4 BAC library using white clover microsatellite markers.

Marker	No. hits in the R3R4 BAC library
ATS123	9
PRS256	5
ATS176	5
ATS055	7
DK501R	7
DNABP	1
RNAH	9
5J9L	5
BE187590	3
Average	**5.7**

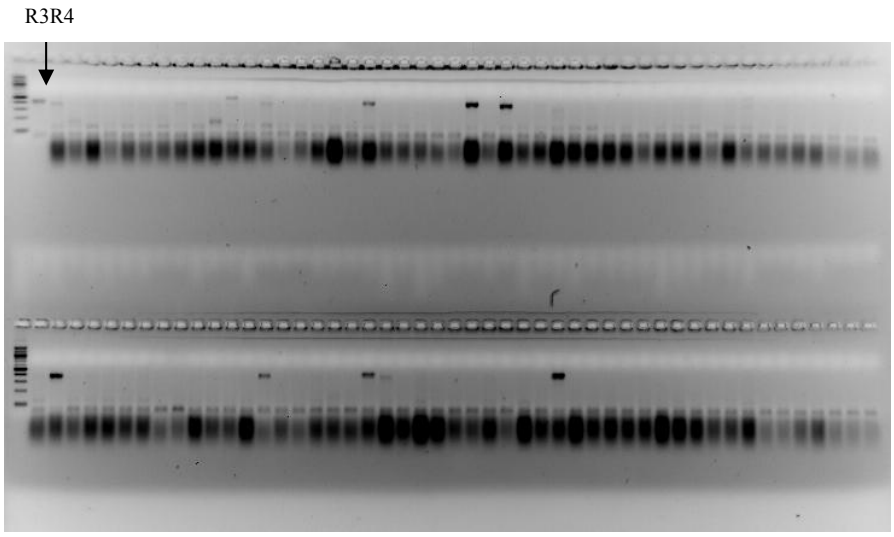


Figure 3.8. BAC library screening. 3% agarose gel showing the analysis of the plate pools with the primer DK501R. Parental line R3R4 was used as a positive control.

3.3.3.3 Estimation of chloroplast contamination

Further characterisation of the R3R4 library was carried out in order to estimate the level of chloroplast contamination. To determine the percentage of BAC clones containing chloroplast DNA sequences, two of the BAC plates were screened against consensus chloroplast (Cc) primer pairs (Chung & Staub, 2003). A multiplex PCR with 3 Cc primers from different regions of the chloroplast genome was carried out on the 2 BAC plates, using DNA from the R3R4 white clover parent as positive control (Figure 3.9). Out of 766 BAC clones tested, 4 showed amplification with one or more Cc primers, thus estimating that less than 0.5% of the library clones carry chloroplast sequences.

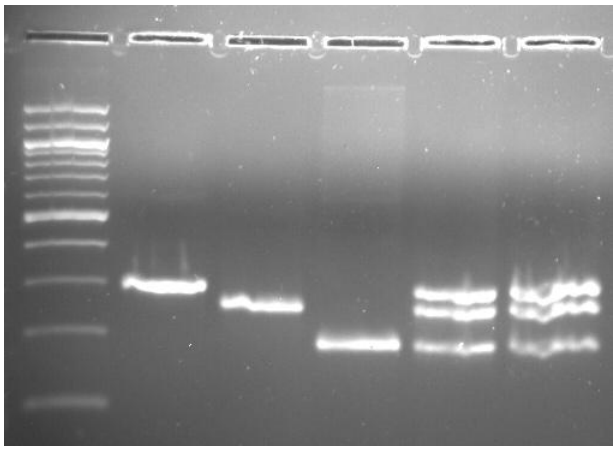


Figure 3.9. Multiplex PCR with three CcSSR primers. Lane 1: 100 bp ladder, Lane 2: CcSSR6, Lane 3: CcSSR15, Lane 4: CcSSR22, Lane 5-6: three primers combined.

3.3.4 BAC-end sequencing

From the previously stated results, we decided to further characterise our BAC library and to test whether a BAC-end sequencing approach might be effective to potentially 'tile' the genome of an uncharacterised species, white clover, on to that of a sequenced species (using *M. truncatula* as the reference species).

The contents of 2x384 well plates were end sequenced, resulting in 1474 BAC-end sequences (BESs) and a total of 725 BES pairs. The average vector-trimmed read length of the BESs was 800 bp with a total read length of 1.16 Mb (Table 3.13). All sequences have been submitted to GenBank, with the accession number ED549329 – ED550789 (Febrer *et al.*, 2007).

Table 3.13. Profile of BAC-end sequences in white clover.

Number of BAC clones sequenced	768
Number of BAC-end sequences	1474
Number of paired BAC-ends	725
Average read length (bases)	800
Total sequence length (Mb)	1.16
Sequence composition	
Protein-coding regions (%)	24.9
Transposable elements (%)	8.5
Microsatellites (%)	2.9

Comparison of the 1474 BES with the TIGR Plant Gene Indices and the TIGR non-identical amino acids database revealed 368 of the sequences (24.9%) could be identified as 'genic' in nature by virtue of good matches to either ESTs (BLASTn, cut-off 1e-20) or representative protein-coding sequences from the non-redundant amino acid database (BLASTx, cut-off 1e-10) (Table 3.14). Of the 368 genic sequences, the top BLAST match in 258 cases was to a legume EST, with 69.4% (179) of the top hits to *Medicago* ESTs, 22.1% (57) to soybean ESTs and 8.5% (22) to *Lotus. japonicus* ESTs (Table 3.14)**.** Overall, 251 of the 368 genic sequences (68.2%) had a good BLAST match (above the cut-off of 1e-10) to a *Medicago* EST.

Of the 1474 BAC-end sequences analysed, 126 (8.5%) were found to contain sequences homologous to transposable elements (E=1e-10) (Table 3.13). The majority of transposable elements belonged to the Ty3_copia type (68.3%) followed by the Ty1_gypsy (23.8%) and LINE (7.9%) types of retrotransposons. No BES was

found to have sequence homology with the CACTA, hAT or monkey type of transposable elements. In comparison, 16.6% of *Medicago* BAC ends for which sequences are available are homologous to transposable elements (E=1e-10). We do not yet know, however, to what extent the higher proportion of transposable elements identified in *Medicago* BAC ends (16.6% vs. 8.5% in clover) is affected by the presence of large numbers of *Medicago* transposable elements in the target database.

Table 3.14. Summary of the comparison of the BAC-end sequences with the TIGR Plant Gene Indices and the TIGR non-identical amino acids database.

Total BES		1474
EST Hit		368
Top Hits		258
Legumes	*Medicago*	179
	Soybean	57
	Lotus japonicus	22
Other	*Arabidopsis*	26
	Rice	25
	Grape	10
	Cotton	6
	C. reinhardtii	6
	S. officinarum	6
	Tomato	5
	Potato	4
	Tobacco	4
	Aquilegia	4
	Maize	3
	Barley	3
	Sorghum	2
	Poplar	2
	Lettuce	2
	Beet	1
	Pepper	1

When the 1474 white clover BESs were compared to *M. truncatula* genome sequence 57% (844) of the BESs had a significant hit (E=1e-10) to *M. truncatula* BACs. Comparisons were made to all *M. truncatula* (Phase I, II and III) BAC sequence available at the time of writing. Of the 1474 BAC-end sequences, 1450 were paired sequences, where the sequence at the other end of the BAC was also known (giving a total of 725 BES pairs). Amongst the 725 BES pairs, 204 had a significant BLAST match (both members of the pair) to *M. truncatula* genome sequence. Of these, a total of 16 (7.8%) were shown to have the equivalent pairs of *M. truncatula* sequence on the same *M. truncatula* BAC clone or contig within a span of 25 Kb to 200 Kb. Clover BACs which fulfil these criteria are putative comparative-tile-BACs, and potentially represent regions of highly conserved gene

content and organisation between clover and *M. truncatula*. The *Medicago* matches to the 16 paired clover BAC-ends ranged in distance from 34.8 Kb to 121.7 Kb with an average of 59.2 Kb (Standard deviation= 35.8 Kb) (Table 3.15). The size of the clover BACs (and thus the distance between the end sequences) was compared to the span by which the paired matches are separated in the *Medicago* genome (Table 3.15). In 11 of the 14 cases, the separation span in clover exceeded that in *Medicago* (with differences ranging from 18.8 Kb to 70.9 Kb), and in three cases the span in *Medicago* exceeded that observed in clover (with differences ranging from 27.9 Kb to 64.6 Kb).

Table 3.15. List of the 16 paired-BAC ends with hits in the same contig/BAC as *Medicago truncatula* and the size of the BAC clones in the two species.

BES	Plate address	Size in white clover (Kb)	Size in *Medicago* (Kb)
WCBE306	27B12	72	51.8
WCBE214	27D03	86	40
WCBE144	27G16	104	67
WCBE053	27I09	93	49.6
WCBE349	27J02	82	115.4
WCBE166	27K12	114	51.6
WCBE368	27L16	100	81.2
WCBE181	27M18	57	121.6
WCBE088	27O07	89	46.3
WCBE614	28D03	95	40.4
WCBE732	28F16	94	23.1
WCBE735	28F22	114	92.8
WCBE546	28G20	75	102.9
WCBE448	28G23	91	3.7
WCBE748	28H24	107	55.6
WCBE666	28L11	-	3.5

3.3.5 Development and mapping of microsatellites from the BAC-end sequences

Amongst the BAC-end sequences, 43 putatively novel microsatellites of a length sufficient for potential marker development were identified (Table 3.7). The predominant motif was the di-nucleotide repeat (58.12%), followed by tri- (20.9%), tetra- (11.6%), mono- (2.3%) and 3 (7%) were compound microsatellites. The number of repeats of the microsatellite motifs ranged from 5 to 34. It was possible to design primers to 34 of these SSRs (Table 3.8). All of these primers were tested in a small set of clover germplasm comprising the R3R4 genotype from which the library was constructed, another genotype referred to as S1S4, and a number of F_1 progeny from a mapping population of which these two genotypes are the parents (Figure 3.10). All of the primers, which were tested successfully, amplified products of the

expected size. Of the 34 primer pairs tested, 21 (61%) revealed a polymorphic pattern that was reliably scoreable over the entire mapping population, which is in keeping with levels of SSR polymorphism previously observed in this population (Chapter 2, Table 2.10). On analysis in JoinMap® 3.0, 18 of the SSRs were successfully incorporated into the genetic map described in Chapter 2. The SSRs are listed in Table 3.16, along with the clover linkage groups to which they are linked and the accession number of the BAC end from which they are derived.

Table 3.16. List of the BAC-end sequences mapped on the white clover genetic linkage map. The accession number of the BAC-end and the map location are also indicated in this table.

SSR ID	Accession number of BAC-end	White clover map location*
WCBE005TF	ED549336	D
WCBE073TF	ED549466	A
WCBE113TF	ED549531	A
WCBE158TR	ED549619	G
WCBE160TF	ED549622	F
WCBE229TR2	ED549750	?
WCBE307TF	ED549890	G
WCBE364TR	ED549998	A, C
WCBE513TF	ED550264	?
WCBE517TRB	ED550273	D
WCBE566TRB	ED550364	D
WCBE578TF	ED550387	F
WCBE602TF	ED550427	G
WCBE655TF	ED550529	D
WCBE692TF	ED550597	G
WCBE712TRB	ED550628	B
WCBE791TRB	ED550780	E

* Chromosomal location according to the naming convention of Barrett *et al*. 2004. Homoeologous groups (eg A1, A2) have not been distinguished.

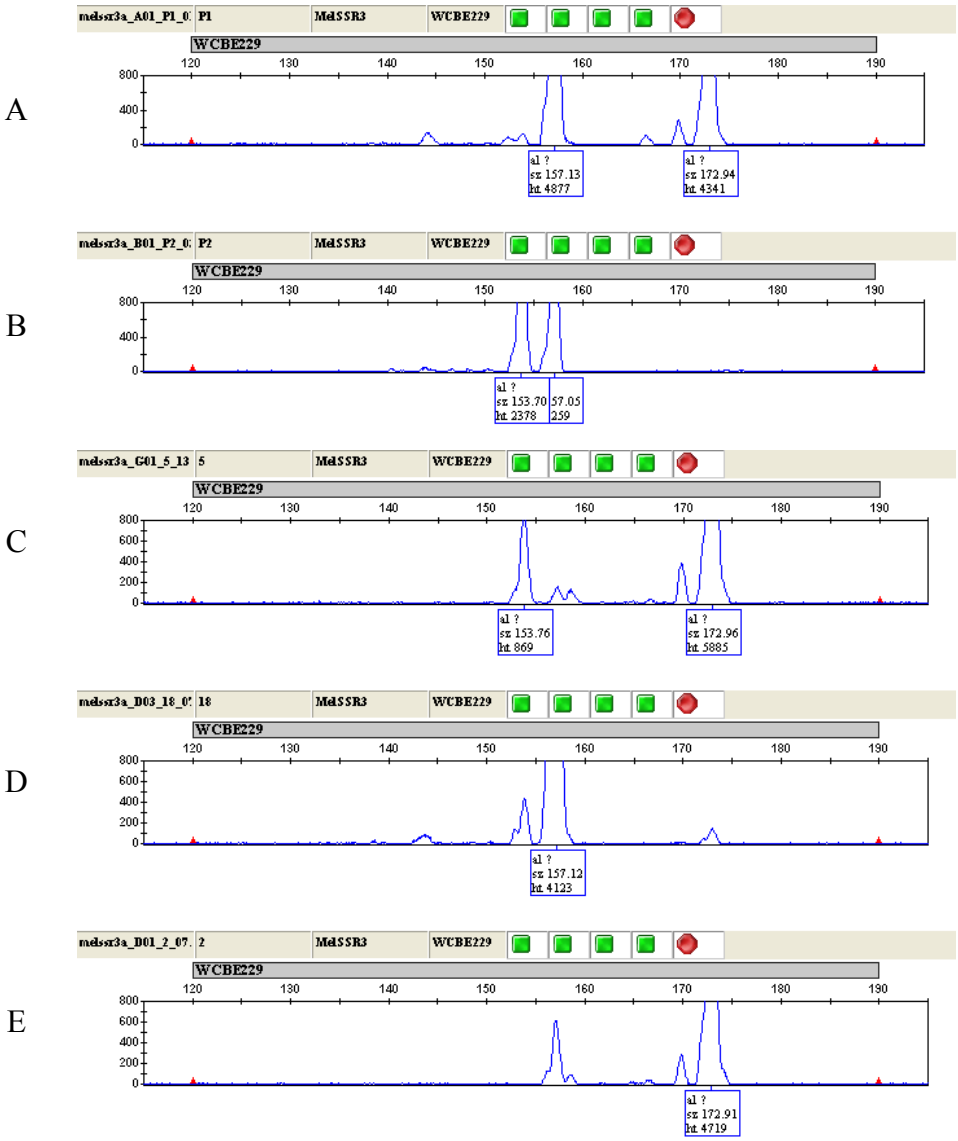


Figure 3.10 An ABI Chromatogram that shows the allelic pattern of a BES SSR (WCBE229) on the parental lines (A, B) and 3 progeny individuals (C, D, E).

3.4. Discussion

The main goal of this part of the study was to carry out a preliminary comparative sequence analysis between white clover and the model legume *Medicago truncatula*.

The first approach used for comparing white clover with *M. truncatula* was a method employing the use of PCR-based markers previously developed by Choi *et al*. (2004a). A set of 95 of those markers were tested on the white clover mapping parents, showing a relatively high level of amplification (73%) from one species to the other, thus indicating how related the two species are. However, only 22 markers amplified in white clover as single amplicons, demonstrating the increased complexity of the white clover genome relative to that of *M. truncatula*. Furthermore, the 22 amplicons were cloned, sequenced and their sequences were compared with the equivalent sequence in *M. truncatula*. The comparison of the sequences displayed some good similarity, which was dependent on the *M. truncatula* marker type. Markers derived from BAC-end sequences seemed to explain a higher similarity to white clover than markers derived from EST sequences. I postulate that the discontinuous nature of the similarity in EST-based markers may be due to differences in the sequence content and position of introns in the two species, although it was not possible to make a comprehensive analysis of this phenomenon within the timescale of this study.

Conclusions from the study by Choi *et al*. (2004a) suggested that a high degree of conservation in gene content and order between the genomes of diploid *M. sativa* (alfalfa) and *M. truncatula* was observed and that many of gene-based genetic markers developed in this study will have applications for comparative genetic mapping in other related legume species. The high level of amplification of these markers in clover certainly suggests that they might be a good source of comparative genetic markers for this species. However, because of ongoing studies in other groups concerning comparative genetic mapping of white clover and *Medicago*, we decided not to pursue this avenue of investigation.

Having established the sequence-based similarity of white clover and *M. truncatula*, we wished to gain a preliminary understanding of the extent of fine-scale conservation of gene-order between the two species. Therefore a second approach

was undertaken with the construction of a three-genome equivalent BAC library of white clover mapping parent R3R4. The library consists of 37,248 clones with an average insert size of approximately 85 Kb. Taking into account that 0.5% of the library clones contain chloroplast DNA, the library comprises a total of approximately 3000 Mb of nuclear DNA. This equates to a 95.8% probability of containing any sequence in the white clover genome (Clarke & Carbon, 1976). The method used for nuclei extraction in this study, while similar in principle to methods used previously to construct BAC libraries, was adapted from a method previously used for the extraction of crude nuclear proteins in gel mobility shift assays, and to our knowledge, this is the first reported use of this method for BAC library construction. It was found to be very successful in white clover in terms of the small amount (10 g) of leaf material needed to produce sufficient high quality HMW DNA for the construction of a 3X library, with reasonably low levels (<1%) of chloroplast contamination. Screening of the BAC library with both white clover microsatellites and PCR-based markers from *Medicago truncatula* resulted in the detection of an average of ~6 clones per marker, consistent with the calculated genome coverage of the library representing three haploid-genome equivalents.

The analysis of BAC clones by *Not*I digestion followed by PFGE showed that the majority of white clover DNA inserts were present as single *Not*I fragment inserts, implying that the white clover genome apparently contains few *Not*I sites, reflecting the low level of GC richness which is a typical feature of dicot genomes in comparison to monocots (Choi *et al.*, 1995; Danesh *et al.*, 1998; Tomkins *et al.*, 1999; Meksem *et al.*, 2000). All of the 86 randomly chosen clones tested contained an insert, implying a low percentage of empty clones in the library.

A comparative sequence analysis was carried out on BAC clones from the BAC library. Comparison of the clover BES with both EST database entries, and the available genome sequence data for *M. truncatula* lends further support to the proposed high level of sequence-based similarity between these two legume species. Out of the 368 clover BES that were classed as genic in nature, ~70% had a significant BLASTn match to *Medicago* ESTs, a figure that is in keeping with the rate of cross-species amplification of *M. truncatula*-derived PCR-based markers in clover mentioned earlier. When compared to *M. truncatula* genome sequence, 57%

(844) of the 1474 clover BES had a good BLAST match. Taking into account potentially spurious matches involving repeat elements, ~50% of the clover end sequences had a good BLAST match in *Medicago* genome sequence. Given the fact that the comparison is based on approximately 168 Mb of *Medicago* sequence contig data, which corresponds to ~34% of the entire genome of *M. truncatula*, the proportion of clover BACs with good *Medicago* matches is actually relatively high. In addition, the high levels of similarity extend beyond those clover BES which we defined as genic by virtue of either a protein or EST match, as 649 of the clover BES which had a good match in *Medicago truncatula* were neither genic (as defined above) nor repeat element-based.

A major goal in comparing our clover BES data to the *Medicago* genome sequence data was to gain a preliminary insight into the extent of microsynteny between the two species, and to investigate the utility of a large-scale BES strategy to 'tile' a significant proportion of the genome of white clover on to that of *Medicago*. The latter would constitute a useful translational genomics platform for gene isolation in white clover in the absence of significant amounts of publicly available sequence information in this species. To that end, we identified all clover BACS which had a significant BLASTn hit to *Medicago* genome sequence at both ends, and the subset of these BACs for which both BAC-ends had good BLASTn matches in *Medicago* genome sequence at a distance consistent with the span of a BAC. Of 204 clover BACs which had a *Medicago* genome sequence match on both ends (at a threshold of 1e-10), 14 had matches which were separated by a compatible distance in *Medicago,* and, consistent with other studies of this nature, we propose that these represent regions of conserved microsynteny between white clover and *Medicago*. It is interesting to note that, in actual fact, the majority (190) of the clover BACs with a BLASTn match to *Medicago* on both ends did not conform to our conditions for identity as comparative tile BACs. It is not possible to comment extensively on the potential microsyntenic relationships between these clover BACs and the *Medicago* genome. Plant genomes are highly repetitive and, in evolutionary terms, subject to segmental duplication events, followed by differential gene loss, processes which tend to result in a fragmentary pattern of conserved microsynteny, even between relatively closely related plant genomes. An additional complication in this case is the fact that white clover is thought to be an allotetraploid descended from a

proposed ancestral hybridisation between two diploid progenitor species, possibly *T. nigrescens* and *T. uniflorum*, with additional introgression of alleles from other diploid species (Ellison *et al.*, 2006). While genetic mapping studies support conservation of macrosyntenic relationships between the two homoeologous genomes of the species, little is known about the extent of conserved microsynteny between these genomes. Both of these factors could contribute to the observed effect, where a significant number of paired-clover BESs seem to "map" to completely different parts of the *Medicago* genome. It would be possible to counteract this by applying more stringent parameters (for the identification of putative orthologues between clover BESs and *Medicago* genome sequence). For example, applying a set of filtering criteria to the current dataset, including a BLASTn threshold of 1e-30, a minimum percent identity of 80%, and a total number of *Medicago* matches lower than 30, results in a drop in the initial number of match pairs from 204 to 22 (data not shown). Such increased stringency filtering might be quite useful in the context of a high-throughput BES initiative, as described below.

Although the proportion of comparative-tile BACs discovered above seems quite low, it does suggest that high throughput BAC-end sequencing of white clover would allow anchoring of a significant portion of the clover genome on to the *M. truncatula* genome sequence. For example, end sequencing 75,000 BACs (equivalent to 6 haploid genome equivalents in a library with our relatively modest insert size) would result in approximately 1400 comparative tile-BACs for white clover, corresponding to 119 Mb of clover genome, not taking into account potential overlapping coverage of the *Medicago* genome by clover comparative-tile BACs. This figure would significantly increase following the sequencing of the remainder of the genespace of *M. truncatula*, (of which approximately an estimated 50-60 percent is currently sequenced; Young and Cannon personal communication). Thus, it is feasible that a significant fraction of the genespace of white clover could be anchored to the genespace of *M. truncatula* using this approach.

We discovered 43 novel clover SSRs amongst the BESs, and explored their utility as genetic markers in a clover mapping population. In a preliminary study, George *et al*. (2006) have proposed that each of the eight homoeologous pairs of clover chromosomes (named A_1, A_2 – H_1, H_2) are broadly homologous to one of the eight

Medicago chromosomes (numbered 1-8). This might lead to the expectation that at least some of the SSR markers derived from clover BES that have significant BLASTn matches to *Medicago* genome sequence would map to the clover linkage group that is the proposed orthologue of the *Medicago* chromosome on which the top BLAST match for that clover BES was found. Such markers could act to anchor the clover genetic map directly to the *Medicago* genome sequence map. This scenario was not found for any of the eight mapped clover SSRs in this study which were derived from BES with a significant BLAST match in *Medicago* genome sequence. However, it is worth noting that six of the eight clover BAC end sequences in question actually had multiple BLAST matches (ranging from 4 to 163 matches at the 1e-10 threshold) to *Medicago* genome sequence, and in four of these cases, lower significance BLAST matches were located on the *Medicago* chromosome consistent with the proposed orthologous group in clover. These results suggest that, while BES-derived microsatellites are a viable source of markers for white clover, it would be problematic to use them in the way suggested above to anchor the genome of clover on to that of *M. truncatula*.

In conclusion, following a pilot experiment suggesting a high level of sequence similarity between *Medicago truncatula* and white clover, we have developed a BAC library of white clover, and preliminary BAC-end sequencing analysis supports significant retention of synteny between the genomes of white clover and *Medicago truncatula*, consistent with other findings of extensive synteny across the cool-season legumes (Choi *et al.*, 2004a; Choi *et al.*, 2004b; Cannon *et al.*, 2006). The BAC library is available as a resource for translational genomics and gene isolation in white clover.

4.0 Comparative genomic studies between white clover and *Medicago truncatula*

4.1 Introduction

In addition to a high degree of conservation of individual gene sequences throughout the plant kingdom, comparative genomics has revealed a high degree of conservation in genome structure, or synteny, among closely related taxonomic groups. In comparative genomics, synteny is often used as a synonym for colinearity and refers to some degree of conservation of gene content, order, and orientation between chromosomes of different species or between non-homologous chromosomes of a single species. For example, gene content appears to be highly conserved with a remarkable degree of colinearity among the grasses, including the grain crops rice, wheat, maize, barley, sorghum, and millet. This is despite large differences in genome size (e.g., 430 Mbp in rice compared with 16,000 Mbp in wheat), which appear to be attributable principally to differences in the amounts of repetitive DNA (associated mainly with retroelements) in intergenic regions and polyploidy (Bennetzen & Freeling, 1997).

While genomic associations within legumes are less well characterised, a number of studies have begun to reveal extensive synteny between the members of this important plant family. Genome conservation was discovered among Phasoloid species, including mungbean (*Vigna radiata*) and cowpea (*V. unguiculata*), extending as long as entire chromosome (Menanciohautea *et al.*, 1993). Macrosynteny between *Medicago truncatula*, *Lotus japonicus*, and four other legume species was examined in detail (Choi *et al.*, 2004c). The results indicate considerable genome-wide synteny between *Mt* and Galegoid legumes (such as alfalfa, pea [*Pisum sativum*], chickpea [*Cicer arietinum*], and *Lj*). The high level of macrosynteny between *Mt* and pea is important, as the pea genome is roughly 10 times larger than *Mt*. Microsynteny between *Mt* or *Lj* and crop species like alfalfa and pea has already enabled the positional cloning of symbiosis genes (Endre *et al.*, 2002; Stracke *et al.*, 2002). A more recent study also showed synteny comparison between *Mt* and *Lt*, where 67% of the *Mt* genome and 64% of the *Lt* genome displayed synteny blocks with extensive conservation of gene order and content (Cannon *et al.*, 2006).

The investigation of microsynteny requires sequencing and annotation of genomic DNA, enabling direct comparison of the sequences using various computational tools. In cases where the level of synteny in a region of interest is high, a potential

inter-specific cloning strategy can be envisaged, and in fact, apart from the many insights comparative sequencing studies can give us into genome evolution and speciation, the use of model species in this way could be viewed as the principal reason for these sorts of analyses. This approach would have greatest utility when the genome of the target species has a relatively large and/or complicated genome, and the genome of the reference species is relatively simple. Thus, the soon to be completed *Medicago truncatula* genome sequence and growing lists of other genomic resources available for this model legume have the potential to be of incredible benefit to genome-based research for crop legumes, and in particular for white clover.

The aim of this chapter was to use the results obtained with the BAC-end sequence analysis in the previous chapter in order to assess the levels of fine-scale conservation of gene order between white clover and *Medicago truncatula* and to validate the idea of large-scale BAC-end sequences for comparative sequencing platform. For this, five clones from the BAC library were sequenced to six-fold coverage and their sequence information was compared to their putatively corresponding *M. truncatula* BAC sequences.

4.2 Material and Methods

4.2.1 Choice of white clover BAC clones

From the BAC-end sequence analysis carried out in Chapter 3, we found that 14 paired BAC-ends were shown to have the equivalent pairs of *M. truncatula* sequence on the same *M. truncatula* BAC clone or contig sequence within a span of 25 Kb to 200 Kb. Of these 14 BES pairs, a subset of five was selected for sequencing to a six-fold level of coverage on the basis of their estimated size (Table 4.1). These 5 clones were picked on the basis that at least one of their end-sequence had a "genic match" (by virtue of either a protein or EST match) in *M. truncatula*. The other criterion for the selection of the white clover BAC clones was their size compared to their corresponding *M. truncatula* sequences.

Table 4.1. List of the 5 paired-BAC ends with hits in the same contig/BAC as *Medicago truncatula,* which were chosen for 6X sequencing.

Plate address	Expected size in white clover (Kb)	Corresponding *M. truncatula* sequence	Estimated span in *M. truncatula* (Kb)
27B12	72	AC146852	51.8
27I09	93	MTCON74	49.6
27K12	114	AC133780	51.6
28F22	114	MTCON5806	92.8
28G20	75	AC152349	102.9

4.2.2 Sequencing of BAC clones

To confirm the sizes of the inserts, the DNA from the five chosen BAC clones was isolated and digested with *Not*I enzyme to release the DNA insert from the cloning vector and analysed by pulse-field gel electrophoresis (PFGE). Each clone was plated into LB agar/Chloramphenicol (12.5 μg/ml) plates and grown overnight at 37^{o}C. A culture stab was then carried into 2 ml tubes containing LB agar/Chloramphenicol (12.5 μg/ml). The 5 tubes containing the 5 BAC clones were sent to GATC Biotech (Konstanz, Germany), where the construction of a shotgun library for each BAC clone (insert size about 2 Kb) and high throughput sequencing of shotgun clones was performed. After removal of the shotgun cloning vector sequence and quality trimming, assembly of the sequence data was performed at GACT using the SeqMan II module of the Lasergene software.

4.2.3 Development of microsatellites from the BAC sequences

From the 5 BACs, simple sequence repeats were identified using the Tandem Repeat Finder software (Benson, 1999). Primers for BAC sequences containing SSRs were designed using the Primer 3 software and the best two primer pairs were selected for each BAC clone (Table 4.2). Each primer pair was tested for amplification by PCR on the genotype used for the BAC library construction (R3R4) and another genotype (S1S4) (Table 4.3 and 4.4).

Table 4.2. List of the SSR primer pairs designed from the BAC sequences.

Name	Motif	Repeat	Forward primer	Reverse primer
27B12.1	tta	11	ACTGGCGCGATACGTTATTT	GACTGAGGAGCCCTCGTATG
27B12.2	at	21	AAACAAAGGTGGTTGATTTGC	GAGATGCAAGCGTGTGTTGT
27I09.1	att	8	GGAAATTAAATGAGGCAATCACA	CGTCACCAACAAAATCATGC
27I09.2	at	13	AAAAACTCAATTTTATTCCCTTTGAA	CCTTTGTGCAATCCTTCTGG
27K12.1	ata	27	ATACAATCAAGCGGGTTTGC	TCCTTTTTCTGATTTGGTTAGAGA
27K12.2	tat	28	CCCACCGCTATTTCAGGTAA	TTGCATTCTCAAGAAGTCAAACA
28F22.1	ac	7	TAGCGGATACACCCGAAAAC	CCCTATCAATTGCTCACACG
28F22.2	tta	10	TTGTTGCTGTTTTTGTACACACC	TGTTGGTGGTTTGGAATTGA
28G20.1	ta	26	AACTAGCGTTGGATGGGTTG	GGCGGCGATGTAATTAAAAG
28G20.2	atg	23	TTGTTCACTGCGCGATTTTA	TGCATCGGTTTGATTTCTTTT

Table 4.3. Standard PCR components.

Component	Volume/single reaction	Final concentration
DNA sample (12.5 ng/µl)	1	1.25 ng
10X Buffer	1	1X
dNTPs (2 mM)	1	200 µM
Forward primer (10 µM)	0.1	0.1 µM
Reverse primer (10 µM)	0.1	0.1 µM
Taq polymerase (5 u/µl)	0.1	0.5 u
dH_2O	6.7	To a final volume of 10 µl

Table 4.4. Amplification condition for standard PCR.

Initial Denaturation	94°C x 3 min	1 cycle
Amplification	94°C x 0.45min	
	50°C x 0.45min	34 cycles
	72°C x 0.45min	
Final extension	72°C x 2min	1 cycle
Hold temperature	4°C	

The BAC-derived SSR primer pairs were then amplified in the mapping population, run on the ABI3100® (Applied Biosystems) and analysed using the ABI Prism® GeneMapper™ Software Version 3.0 (See Chapter 2, Section 2.2.4.2). The segregating SSR markers were mapped onto the genetic linkage map as described in Chapter 2.

4.2.4 Sequence analysis and gene-prediction on BAC sequences

For the purposes of this chapter, the BAC clone sequences were equivalent to Phase I genome sequence data (draft sequence; non-ordered, non-orientated contigs; containing gaps). Where possible, additional information such as the position of the BAC vector in the sequence contigs, and the occurrence of read-pairs from the same subclone in different contigs was used to make preliminary assumptions about the relative order and orientation of contigs. However, due to time and resource constraints, none of these contig orientations were confirmed by further PCR/sequencing.

Sequence contigs arising from the assembly of the BACs were checked for *E. coli* contamination using BLASTn against the *E. coli* database. All *E. coli* derived contigs and all of the contigs below 6 Kb were eliminated for further analysis. The remainder contigs for each BAC and the associated *M. truncatula* sequence(s) were analysed for protein coding genes with the FGENESH Gene structure prediction software using the *Medicago truncatula* prediction matrix (www.softberry.com). The BAC sequences were also compared the *Arabidopsis* protein database using BLASTp and *M. truncatula* ESTs, transposons using BLAST with the NCBI and TIGR databases.

Gene predictions of the white clover BAC clones were also carried out using Genscan by Dr Gregory May at the National Centre for Genomic Resources (SantaFe, NM). However, because the predictions were not performed on the associated *M. truncatula* sequences, the FGENESH Gene prediction was favoured for the rest of the analyses.

Genomic alignments of white clover BAC sequences/contigs and their corresponding *M. truncatula* sequences were performed using the Shuffle-Limited Area Global Alignment of Nucleotides (SLAGAN) algorithm in the VISTA Browser (http://genome.lbl.gov/vista/index.shtml) (Frazer *et al.*, 2004). The SLAGAN algorithm consists of three distinct stages. During the first stage the local alignments between the two sequences are found using the CHAOS tool (Brudno *et al.*, 2003a; Brudno *et al.*, 2003b). Second the maximal scoring subset of the local alignments uder certain gap penalities is picked from a 1-monotonic conservation map. Finally, the local alignments in the conservation map that can be part of a global alignment

are joined into maximal consistent subsegments, which are aligned using the LAGAN global aligner (Brudno *et al.*, 2003b). This software allowed us to compare the two species at the nucleotide level but also to look at the similarities between exons, introns, intron/exon boundaries, intergenic spaces and the rearrangements that occured. In some cases, the reverse complement of *Medicago truncatula* sequences was used for a better visualisation of the similarity between the two species.

4.3 Results

4.3.1 Characteristics of the five BAC clones

The 5 white clover BAC clones were sequenced to a 6-fold coverage and after removal of the shotgun cloning vector sequence and quality trimming, the sequences were assembled into contigs. The number of contigs observed ranged from 1 for BAC 27B12 to 15 for BAC 27K12. However, only the contigs with a sequence length above 6,000 bp were analysed further (Table 4.5). After elimination of these smaller contigs, two BAC clones were composed of single contigs (BAC 27B12 and 28G20) with sequence lengths of 65,105 bp and 66,130 bp respectively, two BAC clones contained two contigs (BAC 27I09 and 28F22) with total sequence lengths of 81,115 bp and 76,375 bp respectively and the remaining BAC clone 27K12 was composed of five contigs with a total sequence length of 106,241 bp. The total G + C content of each clone ranged from 33% for BAC 27I09 to 35.5% for BAC 27B12, with an average of 34% (Table 4.5).

The size of each of the BAC (based on the sum of contig sizes) was slightly lower (by ~10 Kb) than the expected size, with the exception of clone 28F22, where the difference between the expected and the actual size was 38 Kb (Table 4.5).

Table 4.5. The resulting 5 white clover BAC clone sequences after assembly.

Address	Contig	Expected size	Contig size	Total size	G+C content
27B12	1	72 Kb	65,105 bp	65,105 bp	35.5%
27I09	10 9	93 Kb	45,990 bp 31,125 bp	81,115 bp	33%
27K12	18 17 16 15 14	114 Kb	55,487 bp 21,376 bp 16,049 bp 6,787 bp 6,542 bp	106,241 bp	33.4%
28F22	15 14	114 Kb	45,072 bp 31,303 bp	76,375 bp	34.5%
28G20	9	75 Kb	66,130 bp	66,130 bp	33.8%

4.3.2 Genetic mapping of the BACs

In order to obtain a genetic map location in white clover for the BACs, it was decided to develop a number of SSR markers from each BAC. A total of 83 SSRs were identified in the BAC sequences, ranging from 9 SSRs in BAC 28F22 to 24 SSRs in BAC 27K12, corresponding to an overall density of 1 SSR/ 4.7 Kb (212.7 SSRs/ Mb). The SSRs with the highest number of repeats in each BAC were selected and primers were designed using Primer 3. In order to map the position of each of the BACs, two primer pairs for each BAC were chosen for further analysis in the mapping population used in the Chapter 2. These SSRs all amplified in both white clover mapping parents (R3R4 and S1S4) and were then tested on the mapping population. On analysis in JoinMap® 3.0, 8 of the 10 SSRs were successfully incorporated into the genetic map described in Chapter 2. The SSRs are listed in Table 4.6, along with the white clover linkage groups to which they are linked.

Mapping the BAC-derived SSRs provided several useful pieces of information. Firstly, all of the SSRs mapped to single loci, suggesting that there are unlikely to be well conserved paralogous duplications in the clover genome of the genomic intervals represented by these BACs. In addition, where possible, two SSRs per BAC were mapped, and the resulting map locations support the close physical linkage and single locus nature of the SSRs, and the single copy nature of the genomic interval from which they are derived.

As previously mentioned in Chapter 3, George *et al.* (2006) have proposed that each of the eight homoeologous pairs of clover chromosomes (named A_1, A_2 – H_1, H_2) are broadly homologous to one of the eight *Medicago* chromosomes (numbered 1-8). In that chapter, SSRs derived from BAC end sequences (BESs) were mapped in clover. A proportion of these SSRs were derived from clover BESs with highly significant BLAST matches in the *Medicago* genome. Also in that chapter, we explored the possibility that these clover BES-derived SSRs would tend to map to a position/linkage group in clover that was orthologous to the position/linkage group of the best BLAST match in the *Medicago* genome for the clover BES from which the SSR was derived. In fact, this was not the case. In the current study, increased sequencing coverage has allowed us to identify SSRs associated with five of the comparative-tile BACs whose proposed orthologous position in *Medicago* is based

on matches involving both BAC ends. In contrast to the BES-derived SSRs, all of the BAC-contig-derived SSRs map to the clover linkage group that is the proposed homologue of the *Medicago* linkage group on which the equivalent *Medicago* BAC is located.

As already experienced during the course of this and other studies, the SSRs varied in parental and homoeologue specificity. Three SSRs (27B12-2, 27I09-1, 27I09-2) segregated only on one parent and therefore mapped in one parental map (Table 4.6). Similarly, three SSRs (27K12-1, 27K12-2 and 28G20-1) segregated in both parents and mapped on the two parental maps (Table 4.6). However, two SSRs (28F22-1 28G20-2) segregated in both parents but only mapped in one parental map (Table 4.6). One SSR, 27I09-1, which segregated on both parents, mapped on the two homoeologues of linkage group C in the S1S4 parental map.

Table 4.6. List of the BAC sequences mapped on the white clover genetic linkage map.

SSR ID	White clover map location*	Linkage group in R3R4 x S1S4 population
27B12.1	-	-
27B12.2	G	S-14(G)
27I09.1	C	S-5(C), S-6(C)
27I09.2	C	R-6(C)
27K12.1	C	R-5(C), S-5(C)
27K12.2	C	R-5(C), S-5(C)
28F22.1	E	S-12(E)
28F22.2	-	-
28G20.1	D	R-8(D), S-8(D)
28G20.2	D	S-7(D)

*Chromosomal location according to the naming convention of Barrett *et al.* 2004. Homoeologous groups (eg A1, A2) have not been distinguished.

4.3.3 Sequence analysis and gene-prediction

Gene-prediction analysis was performed for the white clover assembled BAC clones and their corresponding *M. truncatula* sequences using the online version of the FGENESH gene prediction software (Table 4.7). Gene prediction of the white clover BAC clones was also performed using GenScan (Section 4.2.4). The two algorithms showed the same prediction on all 5 BAC clones.

Sixteen genes with 53 exons were predicted from the 27B12 BAC clone, with an average gene size of 1710 bp. The gene density in 27B12 was 1 gene/ 3619 bp. Thirteen genes with 71 exons were predicted from the two contigs of 27I09 BAC clone, with an average gene size of 2316 bp. The gene density in 27I09 was 1 gene/ 5378 bp.

Twenty-six genes with 83 exons were predicted from the five contigs of 27K12 BAC clone, with an average gene density of 1 gene/ 3809 bp.

Twenty-one genes with 59 exons were predicted from the two contigs of 28F22 BAC clone, with an average gene size of 1609 bp. The gene density of 28F22 was 1 gene/ 3294 bp.

Seventeen genes with 85 exons were predicted from the 28G20 BAC clone, with an average gene size of 1909 bp. The gene density of 28G20 was 1 gene/ 3466 bp.

Although a significant amount of annotation is already available for the *Medicago* BACs that are the proposed orthologues of the clover BACs described above, for ease of comparison, we performed the same gene prediction analysis as described above on the *Medicago* BACs or contig regions (Table 4.8).

The annotation of both sets of BACs was extended by attempting to further characterise all of the predicted genes by function. This was achieved by screening each predicted gene from the clover BACs and *Medicago* BACs or contig regions against the *Arabidopsis thaliana* protein database using BLASTp, with a significance threshold cut-off value of 1 e-20.

From a total of 108 predicted clover genes, it was possible to functionally classify 45 (38%) of them using this approach. Similarly, for a total of 143 predicted *Medicago* genes, it was possible to identify 66 (46%) of them (Appendix D).

Table 4.7. Features of white clover BAC clones.

BAC clone	27B12	27I09	27K12	28F22	28G20	White clover BACs*
Sequence length (excluding BAC vector)	57,905 bp	69,915 bp	99,041 bp	69,175 bp	58,930 bp	354,966 bp
Total number of genes	16	13	26	21	17	93
Average gene size	1710 bp	2316 bp	1523 bp	1609 bp	1909 bp	1813 bp
Gene density	1 gene/ 3619 bp	1 gene/ 5378 bp	1 gene/ 3809 bp	1 gene/ 3294 bp	1 gene/ 3466 bp	1 gene/ 3816 bb
Exon number	53	71	81	59	85	349
Average number exons per gene	3.3	5.4	3.1	2.8	5.0	3.75
Average exon size	440 bp	247 bp	284 bp	330 bp	208 bp	301 bp
Average number introns per gene	1.9	3.5	1.9	1.6	3.4	2.5
Average intron size	217 bp	329 bp	329 bp	498 bp	283 bp	331 bp

* Based on the total and average of the 5 analysed white clover BACs.

Table 4.8. Features of *M. truncatula* BAC clones or contig regions.

BAC clone or contig region	AC146852	MTCON74*	AC133780	MTCON5806**	AC152349	Mt1.0***
Sequence length	145,314 bp	73,202 bp	127,694 bp	135,422 bp	105,433 bp	251,661,848 bp
Total number of genes	39	16	31	27	31	42,358
Average gene size	1446 bp	2124 bp	1558 bp	2650 bp	1774 bp	1700 bp
Gene density	1 gene/ 3726 bp	1 gene/ 4575 bp	1 gene/ 4119 bp	1 gene/ 5015 bp	1 gene/ 3401 bp	1 gene/ 3703 bp
Exon number	101	68	74	118	121	128,076
Average number exons per gene	2.6	4.25	2.4	4.4	3.9	3.0
Average exon size	261 bp	179 bp	341 bp	222 bp	191 bp	318 bp
Average number introns per gene	1.6	3.25	1.3	3.4	2.9	2.0
Average exon size	449 bp	400 bp	488 bp	497 bp	354 bp	364 bp

* *M. truncatula* MTCON74 contig region from 172,021 bp to 245,221 bp.
** *M. truncatula* MTCON5806 contig region from 29,281 bp to 164,701 bp.
*** Mt1.0 Assembly of the *Medicago truncatula* genome sequence (28/09/2006)

4.3.4 Comparison of the white clover BAC clones with their corresponding regions in *M. truncatula*

The comparison of the white clover BAC clones with their corresponding regions in *Medicago truncatula* was carried out at several different levels. The sequences of both species were first compared at the nucleotide level using the SLAGAN algorithm, and the overall sequence similarity displayed as a dotplot. The sequences were then compared at the gene level using the gene predictions from both species (and accompanying BLASTP results where available). Finally, the graphical output of VISTA allowed a comparison based on sequence similarity, with the gene prediction data overlayed on this representation.

As previously mentioned in Section 4.2.4, the white clover BAC clones sequences were treated as being equivalent to Phase I genome sequence data. For ease of comparison, if the assembled clover BAC clone sequences contained two or more contigs, the order and orientation of the contigs is represented on the basis of the best alignment with the corresponding *M. truncatula* sequence (using the dotplot representation as a guide). If available, supporting data, such as the position of the BAC vector in the contigs, and the presence of read-pairs (ie non-overlapping forward and reverse sequences of BAC subclones) in the appropriate orientations in the ends of contigs, was examined to see if the contig order was supported. To take into account the possibility of local re-arrangements, where possible, initial analysis also involved comparisons with *Medicago* BACs and sequence contigs immediately flanking the proposed orthologue of the clover BAC in question. All comparisons to *M. truncatula* BAC clones or contigs were subsequently confined to a "window" defined by the first and last clover gene-match in the *Medicago* sequence.

4.3.4.1 White clover BAC clone 27B12 vs. *M. truncatula* AC146852

The white clover BAC clone 27B12, which mapped to LG G, corresponds to *M. truncatula* BAC clone AC146852, located on chromosome 5. The sequence of 27B12 was composed of only one contig. The dotplot comparison between 27B12 and AC146852 showed good sequence alignment similarity along the white clover

BAC clone, with ~60 Kb of available clover sequence showing similarity to a similarly sized region of *Medicago* genome sequence (Figure 4.1).

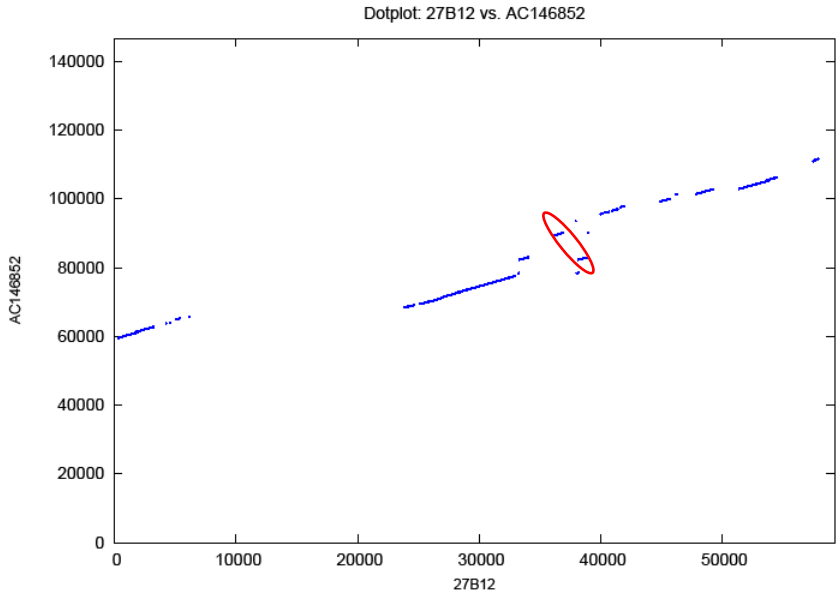


Figure 4.1. Dotplot representing sequence alignment similarity between white clover 27B12 clone and its corresponding *M. truncatula* AC146852 clone. The red circle represents the duplication event observed between 27B12 and *M. truncatula* AC146952.

Comparison of gene sequences showed that 9 of the 19 predicted genes of white clover 27B12 clone displayed correspondence with predicted genes in the *M. truncatula* clone (Figure 4.2, Table 4.9). In all cases, the predicted genes had the same relative transcriptional direction.

Table 4.9. BLASTp results of the analogues genes in white clover 27B12 and *M. truncatula* AC146852.

27B12	AC146852	BLASTp Result
Gene 1	Gene 17	Putative DNA binding protein
Gene 6	Gene 18	Centromere protein
Gene 7	Gene 18	-
Gene 9	Gene 22	N-acetyltransferase
Gene 10	Gene 24	N-acetyltransferase
Gene 11	Gene 26	Structural constituent of ribosome
Gene 12	Gene 27	Unknown protein (NP_18893.1)
Gene 14	Gene 28	Selenium binding protein
Gene 17	Gene 30	T25N20.8

Gene 1 in 27B12 corresponded to gene 17 in AC146852. Genes 6 and 10 in 27B12 were both associated with parts of gene 18 in AC146852. We observed a "reciprocal

duplication" event, where genes 9 and 10 in 27B12 represented apparently duplicated acetyltransferases, which were analogous to genes 22 and 24 in AC146852 (this duplication is apparent on the dotplot). Genes 11, 12, 14 and 17 in 27B12 corresponded to genes 26, 27, 28 and 30 in AC146852.

Overlaying the gene prediction onto the Shuffle Lagan alignment in Figure 4.2 allowed us to visualise the relationship between sequence similarity and gene content in the BAC sequences from clover and *Medicago*. As might be expected, sequence similarity was generally restricted to regions corresponding to predicted genes in the two species, with relatively little sequence homology observable in apparent intergenic regions. Within these regions, similarity is greatest in exonic regions, and variable, but generally lower in intronic regions (Figure 4.3).

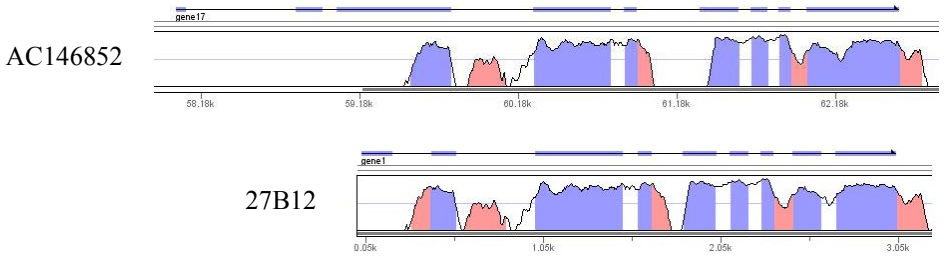


Figure 4.3. Close up of a region of similarity between white clover 27B12 clone and *M. truncatula* AC146852 clones. This shows the good similarity in both exonic and intronic regions.

In the Shuffle Lagan alignment, a limited number of areas of homology occur in apparently intergenic regions (these peaks are shaded pink on the alignment). For instance, between genes 1 and 2 in clover 27B12 there is a region exhibiting similarity to a region between *Medicago* genes 17 and 18. We would hypothesise that these regions are likely to represent genes common to both species that have not successfully been predicted by the gene prediction software. A similar phenomenon was observed for a region between clover genes 11 and 12, which exhibits similarity to a region between *Medicago* genes 26 and 27 (Figure 4.2).

In addition to apparent "intergenic" homology, there are instances in which homology between the two species extends beyond, but is directly contiguous to, the gene predictions in both species. For example a significant region preceding gene 6

in clover 27B12 exhibits similarity to a region preceding gene 18 in *M. truncatula* AC146852. Interestingly, the first predicted exon for both species is "buried" in this sequence. Again, the most likely explanation is that the gene prediction software has incorrectly predicted the extent of the gene in both species. This idea is supported by the fact that, in both the above, and in the previously described cases of apparent intergenic homology, a discontinuous pattern of sequence-homology similar to that conferred by the intron/exon structure of the predicted genes is observed.

A further complication arose from the observation that two genes (gene 6 and 7) in white clover 27B12 were associated with one gene (gene 18) in *M. truncatula* AC146852. Examination of the Shuffle-Lagan alignment revealed that this was not a duplication, but again, a probable miscall by the gene prediction software, resulting in either the erroneous fusion of two genes in *Medicago* or the splitting of a single gene in clover. However, to the fact that gene 7 in white clover did not encode protein, whereas clover gene 6 and *Medicago* gene 18 both encoded centromere proteins (according to the BLASTP analysis) suggests that gene 6 and 7 in white clover really only represent a single gene.

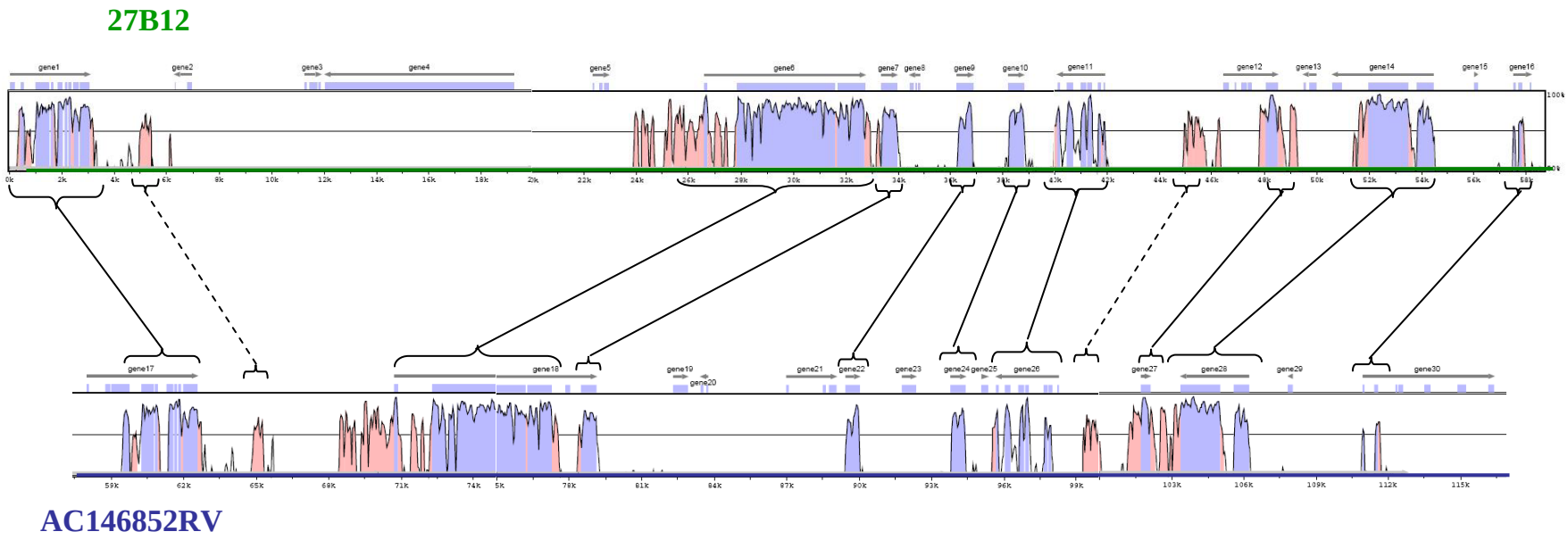


Figure 4.2. Shuffle Lagan sequence alignment between white clover 27B12 clone against *M. truncatula* BAC clone AC146852. Gene predictions of the white clover and *M. truncatula* clones are showed on the top, with the blue boxes representing the exonic regions predicted by FGENESH. The arrows above represent the direction of transcription for each predicted genes. Peaks on the alignment represent areas of similarity between the two species. Where peaks are shaded blue, the area of sequence homology corresponds to a predicted exon. No exon has been predicted for areas of similarity where peaks are shaded pink. Red double arrows, if present, mark the separate contigs from white clover. The dotted lines represent orthologous genes not predicted by FGENESH.

4.3.4.2 White clover BAC clone 27I09 vs. *M. truncatula* sequence MTCON74

The white clover BAC clone 27I09, mapped on LG C, corresponded to a region of *M. truncatula* contig sequence MTCON74, located on chromosome 7. The sequence of 27I09 was composed of two major contigs, which were ordered based on the dotplot view (Figure 4.4). Examination of the assembled clover contigs revealed that the BAC vector was present at the beginning of Contig 1 and the end of Contig 2, supporting the order and orientation represented in Figure 4.4 (read-pair data also supported the order). The *M. truncatula* contig region used for comparison was selected on the basis of the alignment with the clover, giving a window of comparison in *M. truncatula* of 73 Kb compared to 69 Kb in clover. The dotplot alignment of 27I09 and *M. truncatula* MTCON74 contig region showed a good sequence alignment between the two species within this window (Figure 4.4).

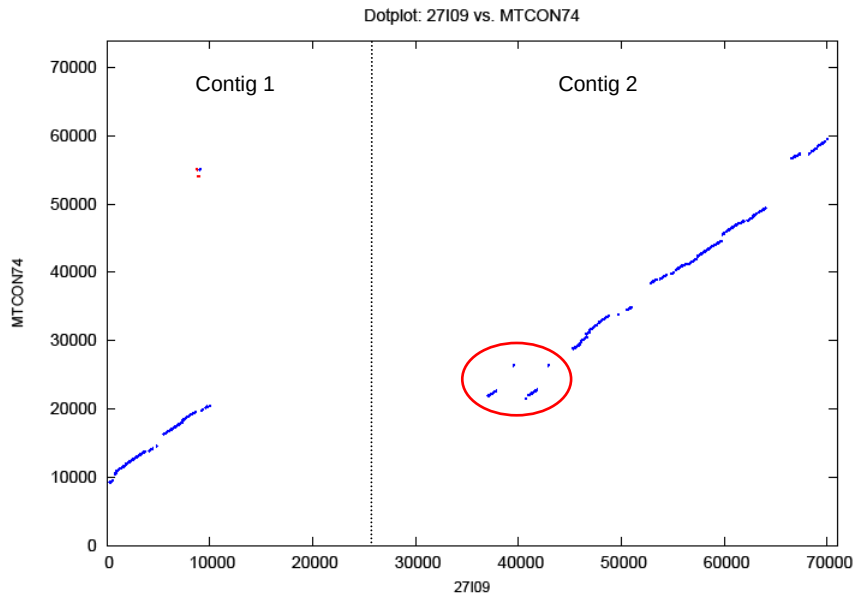


Figure 4.4. Dotplot representing sequence alignment similarity between white clover 27I09 clone and its corresponding *M. truncatula* MTCON74 sequence contig. The red circle represents the duplication event observed between 27I09 and *M. truncatula* MTCON74.

The Shuffle LAGAN alignment, including the gene prediction for the 2 sequences, revealed the conservation of genes between the two species (Figure 4.5). A total of seven genes in white clover BAC 27I09 showed high similarity with six genes in MTCON75 (Table 4.10). As in the previous case, all the predicted genes had the same relative transcriptional direction.

In contrast with the "reciprocal duplication" observed in the previous comparison, an "asymmetric duplication" event was observed in this clone, where gene 8 and gene 9 in 27I09 were similar to gene 4 in MTCON74, both encoding the protein pepsin A (Table 4.10, Figure 4.5). This duplication is also represented in the dotplot (Figure 4.4).

Table 4.10. BLASTp results of the analogues genes in white clover 27I09 and *M. truncatula* MTCON74 contig region.

27I09	MTCON74	BLASTp Result
Gene 1	Gene 2	ATP binding/ kinase/ protein kinase
Gene 2	Gene 3	Double stranded RNA binding
Gene 8	Gene 4	Pepsin A
Gene 9	Gene 4	Pepsin A
Gene 10	Gene 6	Catalytic
Gene 11	Gene 9	Nucleotide binding
Gene 12	Gene11	Unknown protein (NP_187146.2)

As previously, there are instances in which homology between the two species extends beyond, but is directly contiguous to, the gene predictions in both clover and *Medicago*, suggesting that the gene prediction software has incorrectly predicted the extent of the gene in both species. For example a region at the end of gene 11 in clover 27I09 exhibits similarity to a region at the end of gene 9 in *M. truncatula* MTCON74. In the Shuffle LAGAN alignment, we also observed an apparently non-genic region at the end of 27I09 which seems to correspond to gene 12 in *M. truncatula* MTCON74, suggesting that FGENESH may have successfully predicted a gene in *Medicago*, but not in clover, for these sequences.

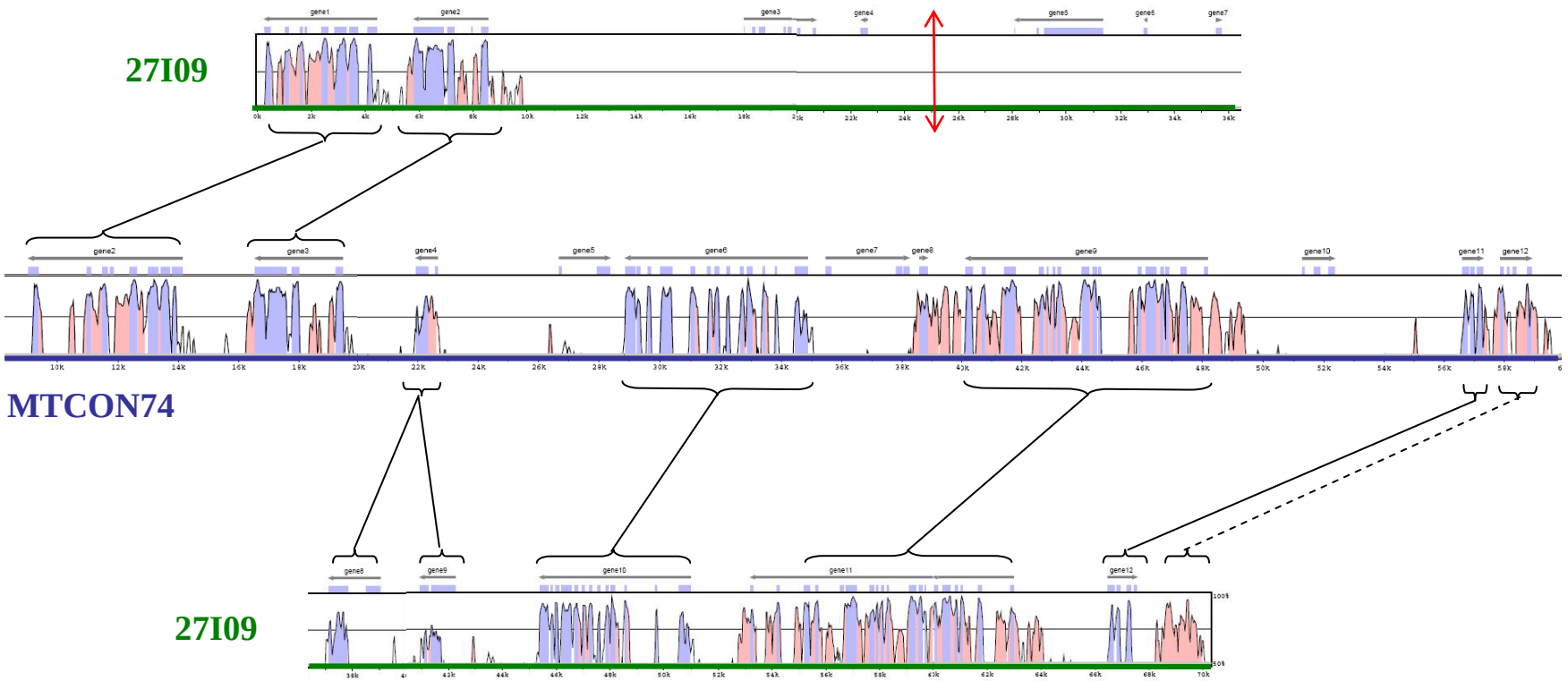


Figure 4.5. Shuffle Lagan sequence alignment between white clover 27I09 clone against *M. truncatula* sequence contig MTCON74. Gene predictions of the white clover and *M. truncatula* clones are showed on the top, with the blue boxes representing the exonic regions predicted by FGENESH. The arrows above represent the direction of transcription for each predicted genes. Peaks on the alignment represent areas of similarity between the two species. Where peaks are shaded blue, the area of sequence homology corresponds to a predicted exon. No exon has been predicted for areas of similarity where peaks are shaded pink. Red double arrows, if present, mark the separate contigs from white clover. The dotted lines represent orthologous genes not predicted by FGENESH.

4.3.4.3 White clover BAC clone 27K12 vs. *M. truncatula* AC133780

The white clover BAC clone 27K12, which maps to LG C, corresponded to *M. truncatula* clone AC133780, located on chromosome 7. The sequence of 27K12 was composed of five contigs (Contig 1 to Contig 5). As in the previous comparison, we attempted to order and orient the clover contigs on the basis of the nucleotide alignment with the *Medicago* sequence. In contrast to the previous example, however, while the dotplot shows a good level of colinearity between 27K12 and AC133780 over the length of Contigs 1, 2 and 3, the similarity between clover contigs 4 and 5 and AC133780 is much lower. Examination of read-pair data for the contigs was inconclusive, with several possible orders supported, however, the BAC vector was found at the beginning of contig 1 and the end of contig 3, suggesting that these are the outermost contigs. Because the low level of sequence similarity between AC133780 and contigs 4 and 5 makes it difficult to suggest a relative order for these sequences based on the alignment, they have arbitrarily been placed as the final two contigs for comparison purposes in Figures 4.6 and 4.7.

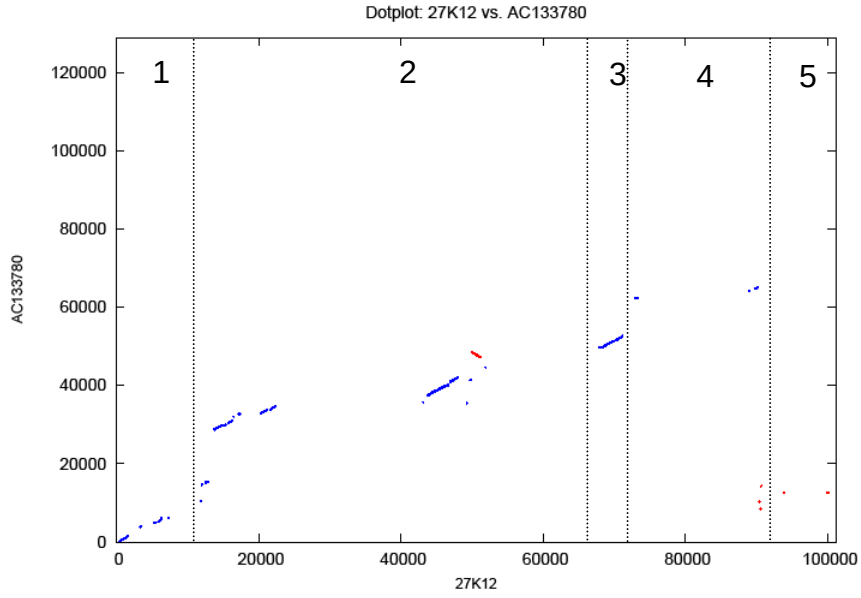


Figure 4.6. Dotplot representing sequence alignment similarity between white clover 27K12 clone and its corresponding *M. truncatula* AC133780 clone. The regions in red represent inversions in the sequence alignment.

With a single exception, all conserved genes in the two species had the same relative transcriptional direction. Gene 10 in clover 27K12 and gene 15 in *Medicago* AC133780 both encode the same predicted protein but are in opposite transcriptional

direction (Figure 4.7). This corresponds to the obvious inversion between AC133780 and contig 2 in the dotplot comparison.

Genes 1, 2, 3 and 4 in 27K12 are similar to Genes 1, 3, 9 and 11 in AC133780. While the four clover genes are apparently consecutive as represented in Figure 4.7, there is an approximately 20 Kb interval, containing 5 predicted genes, between genes 3 and 9 in AC133780. However, in 27K12, genes 1 and 2 are on contig 1, while genes 3 and 4 are on contig 2. Given the previous discussion regarding clover contigs 4 and 5, it is possible that this large apparent discontinuity represents the position of one or both of these sequences in the clover BAC. Examination of the nucleotide similarity between the two species, as represented on the dotplot, certainly shows that there is some limited homology between contigs 4 and 5 and the appropriate region in ACC133780. However, the extent of the similarity relative to comparisons involving the other contigs was low, and it was felt that the proposition of an order and orientation for the contigs on this basis would be somewhat speculative.

As observed in 27B12, there is a "reciprocal duplication" event in this clone, where two genes in 27K12 (genes 3 and 4) both encode for the same protein (a metal ion binding protein), as two genes in AC133780 (genes 9 and 11) (Table 4.11). As in previous comparisons, there are instances of apparent intergenic homology that may represent unpredicted genes common to both species. For example, between genes 1 and 2 in clover 27K12 there is a region exhibiting similarity to a region between *Medicago* genes 1 and 3 (Figure 4.7). This also occured for a region between clover genes 2 and 3, which exhibits similarity to a region between *Medicago* genes 6 and 7, and a region between clover genes 22 and 23, with apparent similarity to a region between *Medicago* genes 16 and 17. In addition, there are instances in which homology between the two species extends beyond, but is directly contiguous to, the gene predictions in both species. For example a significant region preceding gene 9 and gene 16 in clover 27K12 exhibits similarity to a region preceding gene 12 and gene 16, respectively, in *M. truncatula* AC133780.

Table 4.11. BLASTp results of the analogues genes in white clover 27K12 and *M. truncatula* AC133780.

27K12	AC133780	BLASTp Result
Gene 1	Gene 1	SRG3 (Senescence related gene 3)
Gene 2	Gene 3	-
Gene 3	Gene 9	Metal ion binding
Gene 4	Gene 11	Metal ion binding
Gene 9	Gene 12	PBS1 kinase serine/ threonine-protein like
Gene 10	Gene 15	Unknown protein (NP_566505.1)
Gene 16	Gene 16	Transcription factor/ zinc ion binding

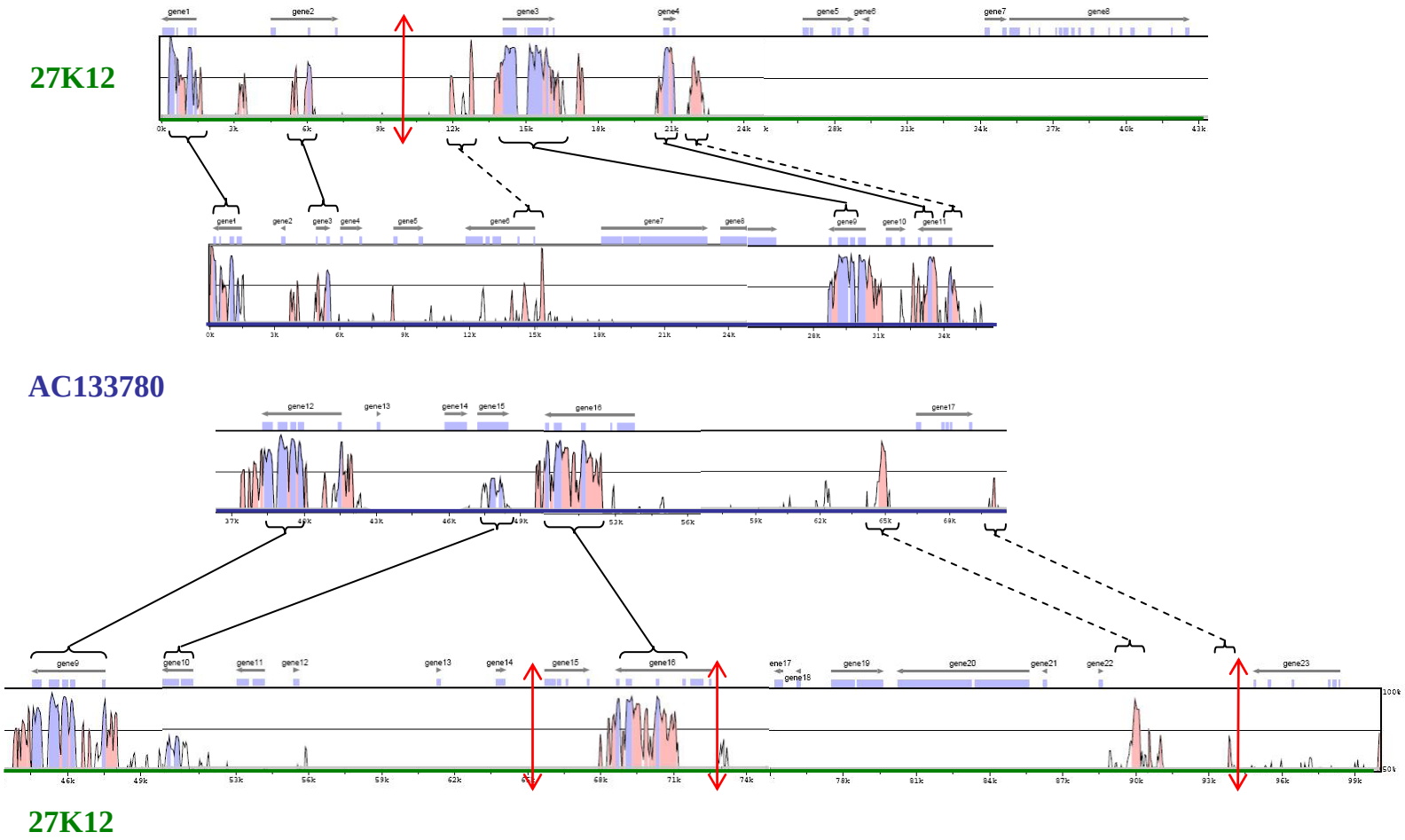


Figure 4.7. Shuffle Lagan sequence alignment between white clover 27K12 clone against *M. truncatula* AC133780. Gene predictions of the white clover and *M. truncatula* clones are showed on the top, with the blue squares represent the exonic regions. The red double arrows correspond to the white clover contig separation. The dotted lines represent orthologous genes not predicted by FGENESH.

4.3.4.4 White clover BAC clone 28F22 vs. *M. truncatula* MTCON5806

The white clover BAC clone 28F22, mapped on LG E, had its correspondence with a region of *M. truncatula* contig MTCON5806, located on chromosome 1. The sequence of 28F22 was composed of two major contigs, which were ordered based on the dotplot view (Figure 4.8). Using the criteria previously outlined, approximately 70 Kb of clover sequence was compared to 90 Kb of *Medicago* sequence. The comparison of 28F22 and the *M. truncatula* sequence contig showed a good colinearity (Figure 4.8).

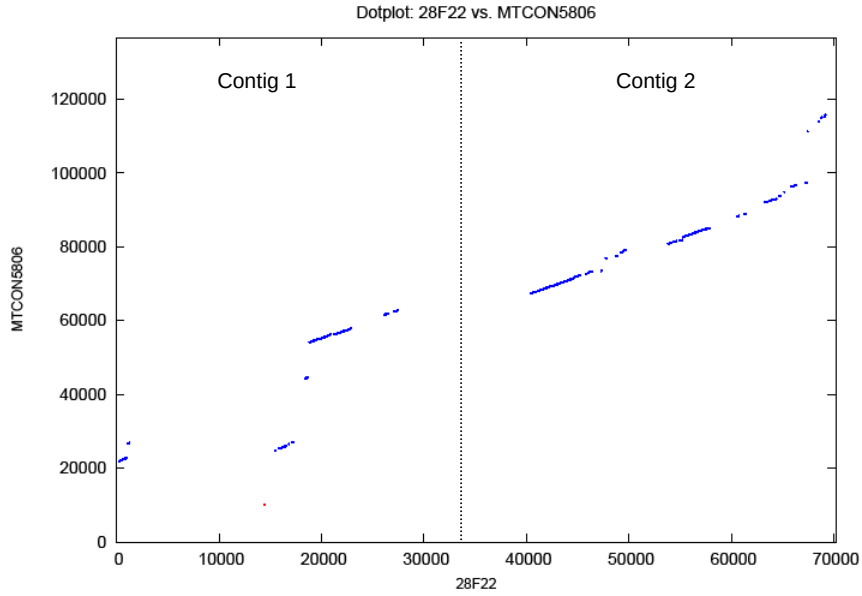


Figure 4.8. Dotplot representing sequence alignment similarity between white clover 28F22 clone and its corresponding *M. truncatula* clones AC131240 and AC146755.

The Shuffle LAGAN alignment also displayed a good conservation of genes between the white clover BAC clones and its corresponding *M. truncatula* sequence (Figure 4.9). Seven genes in 28F22 showed a good conservation with genes with *M. truncatula* MTCON5806 (Figure 4.9, Table 4.12). In all cases, the predicted genes had the same relative transcriptional direction.

As observed in the previous comparisons, there are instances of apparent intergenic similarity. For example, between genes 6 and 7 in clover 28F22 there is a region exhibiting similarity to a region just after *Medicago* genes 11; and also between genes 15 and 16 and genes 17 and 19, there are existing similarity to a region between *Medicago* genes 17 and 19 and genes 19 and 21 respectively.

One gene (gene 20) in white clover 28F22 was associated with two genes (genes 22 and 26) in *M. truncatula* MTCON5806. Examination of the Shuffle-Lagan alignment revealed that the two genes in *Medicago* MTCON5806 are each similar to one part of predicted gene 20 in clover, implying that .this was a probable miscall by the gene prediction software, resulting in the erroneous fusion of two genes in white clover.

Table 4.12. BLASTp results of the analogues genes in white clover 28F22 and *M. truncatula* MTCON5806 contig region.

28F22	MTCON5806	BLASTp Result
Gene 1	Gene 4	F21J9.22
Gene 6	Gene 5	-
Gene 7	Gene 14	Camodulin binding/ triacylglycerol lipase
Gene 9	Gene 15	Structural constituent of ribosome
Gene 15	Gene 17	Transcription factor
Gene 17	Gene 19	D-alanyl-D-alanine endopeptidase
Gene 20	Gene 22	Carboxylic ester hydrolase
Gene 20	Gene 26	Carboxylic ester hydrolase

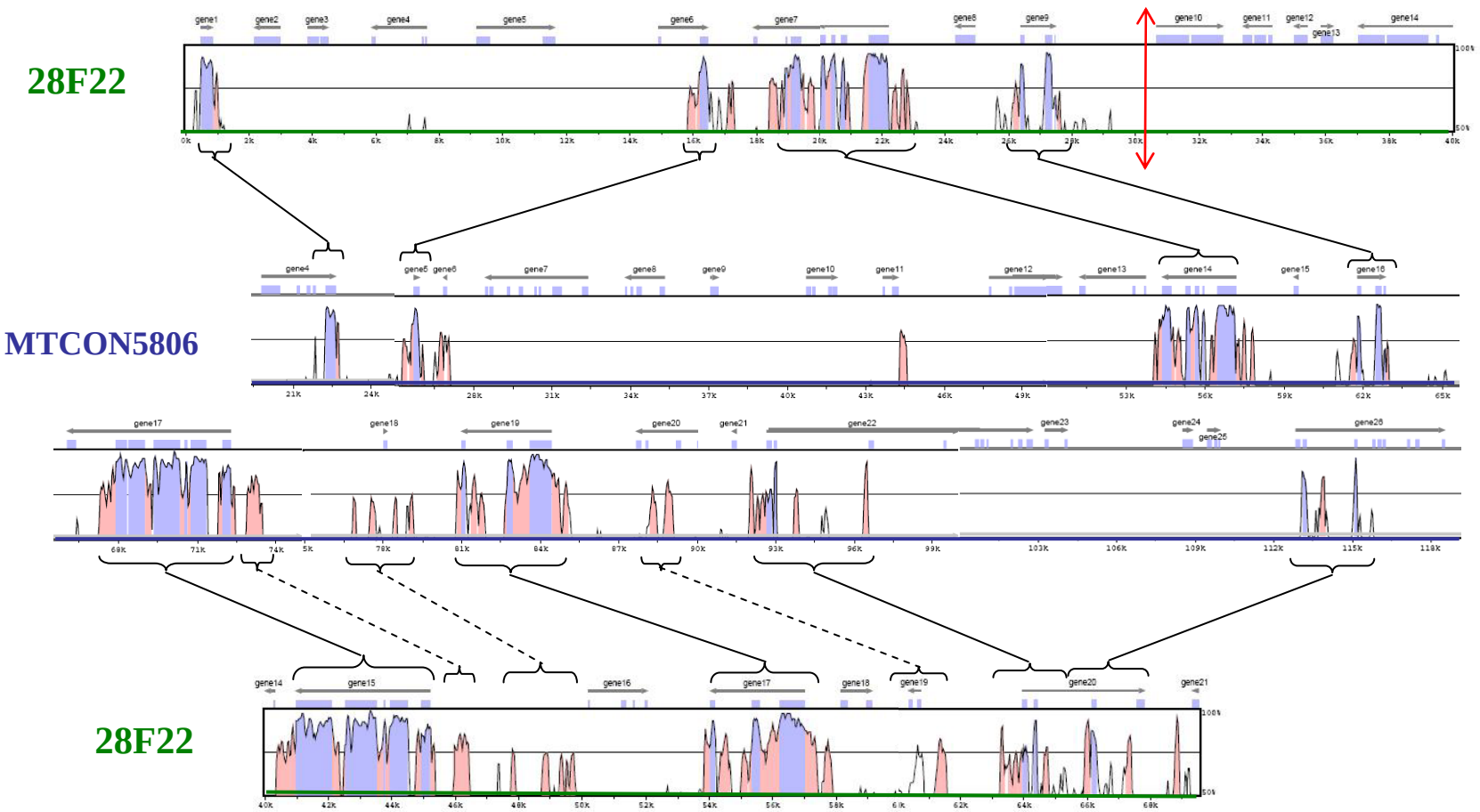


Figure 4.9. Shuffle Lagan sequence alignment between white clover 28F22 clone against *M. truncatula* MTCON5806. Gene predictions of the white clover and *M. truncatula* clones are showed on the top, with the blue squares represent the exonic regions. The red double arrow corresponds to the white clver contig separation. The dotted lines represent orthologous genes not predicted by FGENESH.

4.3.4.5 White clover BAC clone 28G20 vs. *M. truncatula* AC152349

The white clover BAC clone 28G20, mapped on LG D corresponds to *M. truncatula* BAC clone AC152349, located on chromosome 4. The sequence of 28G20 was composed of only one contig. The dotplot comparison between 28G20 and AC152349 showed good sequence alignment similarity along a similarly sized region in both sequences (Figure 4.10).

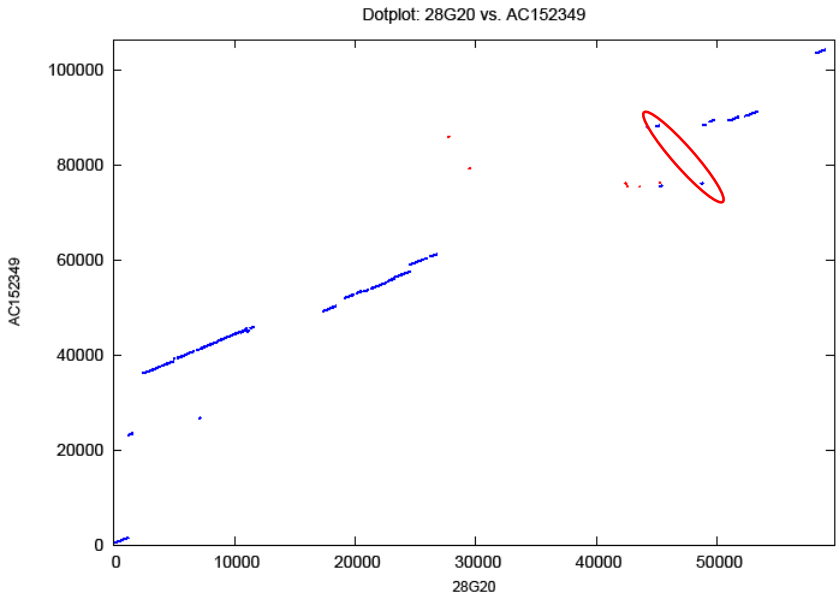


Figure 4.10. Dotplot representing sequence alignment similarity between white clover 28G20 clone and its corresponding *M. truncatula* clone AC152349. The red circle represents the duplication event observed between 28G20 and AC152349.

Conservation of gene sequences showed that 9 genes of white clover 28G20 clone displayed correspondence with the *M. truncatula* clone (Figure 4.11, Table 4.13). In all cases, the predicted genes had the same relative transcriptional direction in both species.

Genes 1, 2 and 3 in 28G20 exhibited conservation with genes 1, 12 and 14 in AC152349. There is a long interval of approximately 30 Kb, between genes 1 and 12 in AC152349, which contains 10 predicted genes, but this does not seem to have a corresponding interval in the clover BAC. Gene 4 in clover 28G20 showed high similarity to two genes in *Medicago* AC152349 (genes 15 and 16). However this is not a duplication event, as the two genes in *Medicago* encoded for two different proteins, disulfide oxidoreductase for gene 15 and glucose inhibited division protein A for gene 16. As mentioned in clover 27B12 and 28F22, this event reflects a

probable miscall by the gene prediction software, resulting in either the erroneous fusion of two genes in *Medicago* or the splitting of a single gene in clover.

As observed in several previous cases, an "asymmetric duplication" event was observed in this comparison, where gene 13 and gene 14 in 28G20 were similar to gene 26 in AC152349, both encoding for the protein YAP169 gibberellin 20 oxidase (Table 4.13, Figure 4.11). This duplication is also represented in the dotplot (Figure 4.10). Genes 15 and 16 in 28G20 showed good conservation with genes 27 and 31 in AC152349.

Table 4.13. BLASTp results of the analogues genes in white clover 28G20 and *M. truncatula* AC152349.

28G20	AC152349	BLASTp Result
Gene 1	Gene 1	BAB11018.1
Gene 2	Gene 12	ATP binding/ ATP dependent helicase/ DNA binding
Gene 3	Gene 14	Inorganic diphosphatase/ Mg2+ ion binding/ pyrophosphatase
Gene 4	Gene 15	Disulfide oxidoreductase
Gene 4	Gene 16	Disulfide oxidoreductase
Gene 13	Gene 26	YAP169 gibberellin 20 oxidase
Gene 14	Gene 26	YAP169 gibberellin 20 oxidase
Gene 15	Gene 27	Transcription factor
Gene 16	Gene 31	Ankyrin-like protein

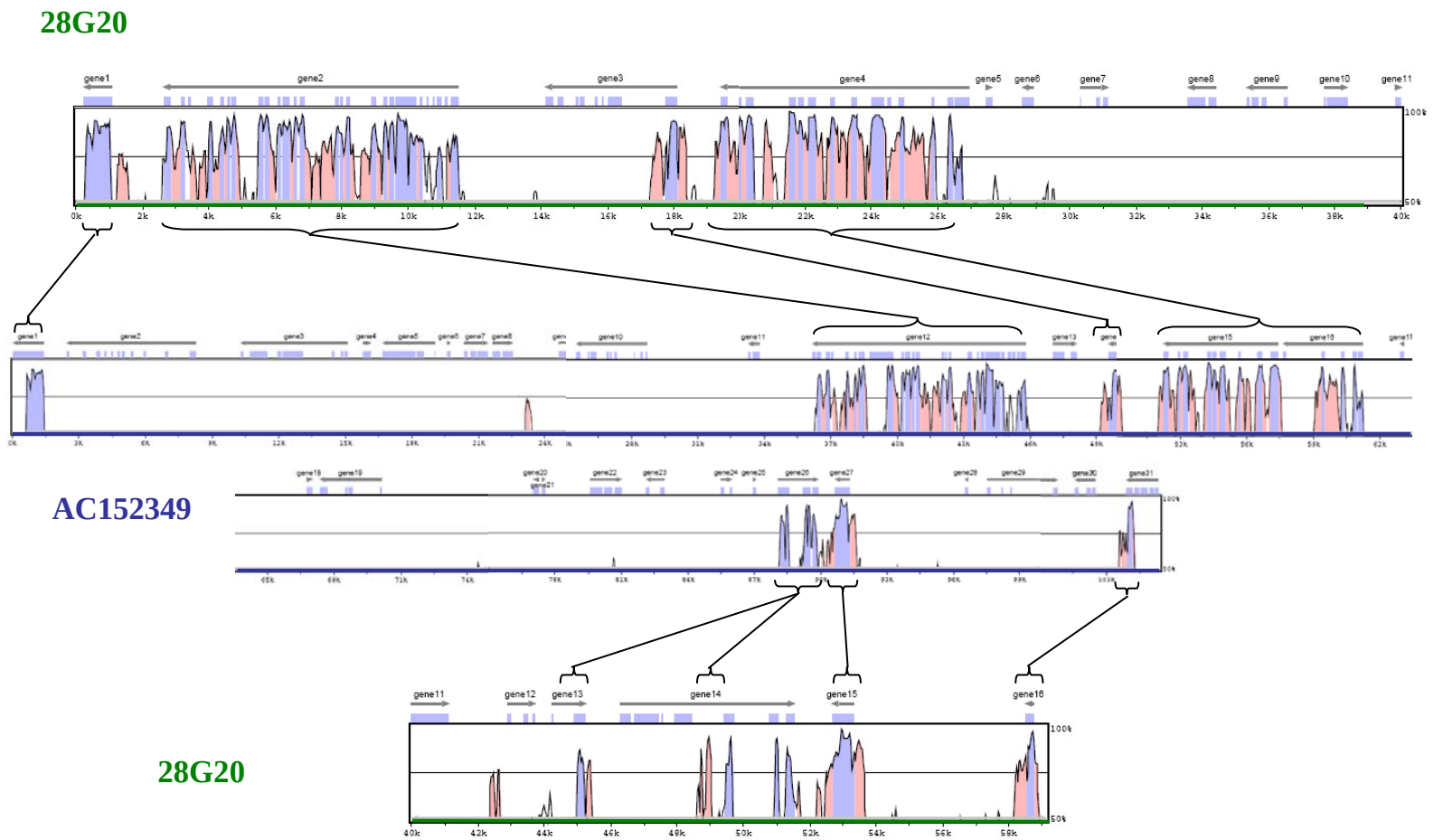


Figure 4.11. Shuffle Lagan sequence alignment between white clover 28F22 clone against *M. truncatula* AC152349 clone. Gene predictions of the white clover and *M. truncatula* clones are showed on the top, with the blue squares represent the exonic regions.

4.3.4.6. Transposable elements

The white clover BAC clones and their corresponding *M. truncatula* sequences were searched for transposable elements (TEs) using the TIGR *Fabaceae* database (E value e-06 and % identity >50%). A total of 46 TEs were found in the white clover BAC clones, ranging from 1 in 27B12 to 19 in 27K12 (Table 4.14). For the *M. truncatula* sequences, a total of 53 TEs were found, ranging from 5 TEs in AC146852 and in MTCON74 to 16 TEs in MTCON5806 (Table 4.14).

Table 4.14. Summary of transposable elements found in the white clover BAC clones and their corresponding *M. truncatula* sequences.

Sequence ID	Number of TEs	Total length (bp)	Average length (bp)
27B12	1	497	497
AC146852	5	2518	503
27I09	16	11461	716
MTCON74	5	2770	554
27K12	19	13365	703
AC133780	12	10143	845
28F22	5	3087	617
MTCON5806	16	9712	607
28G20	5	2401	480
AC152349	15	7669	511

The majority of TEs obtained in the sequences of the two species originated from two types, *M. truncatula* CACTA type (51% in both white clover and *M. truncatula* sequences) and Soybean transposable element *tmg*1 (41% in white clover and 30% in M. *truncatula* sequences). Thus, overall there seems to be no great difference in the repeat-element complement between the two species. Comparison of the relative positions of the TEs in the two species revealed no similarities.

4.4 Discussion

In the previous chapter, we proposed a strategy, based on BAC-end sequencing, to identify regions of the clover and *Medicago truncatula* genomes that exhibit high levels of conserved microsyteny. During the BES analysis, we found that 14 white clover BACs, which had a *Medicago* genome sequence match on both ends which were separated by a compatible distance in *Medicago*. We proposed that the respective clover and *Medicago* BACs are likely to represent orthologous regions in the two genomes, with an expected high degree of gene content and order conservation. During the course of this chapter, a subset of 5 of these 14 BAC clones was subjected to 6-fold coverage sequencing to test this hypothesis.

Annotation of the assembled clover BAC clones immediately demonstrated that they were all gene rich, and thus originated from the euchromatic portion, or the gene-space of white clover. The gene density of the white clover BAC clones (1 gene/ 3.8 Kb) was very similar to overall the gene density of the *Medicago* genespace sequence (1 gene/ 4.4 Kb, based on the Mt1.0 Assembly of the *Medicago truncatula* genome sequence). Other characteristics such as average gene length, GC content, average exon and intron length and number of exons and introns per gene were also quite comparable for the two species.

As an initial test of whether the clover and *Medicago* sequences under examination were likely to be orthologous, a number of SSR markers were developed from each clover BAC, and placed on to the genetic map developed during Chapter 2. In closely related species, orthology extends to the chromosome level. White clover and *Medicago* have the same basic chromosome number (eight, when the fact that clover is allotetraploid is taken into account). As previously mentioned, George *et al*. (2006) have demonstrated that each of the eight homoeologue pairs of white clover chromosomes is broadly orthologous to one of the eight *Medicago* chromosomes (the relationships are: Mt 1 = Tr E, Mt 2 = Tr F, Mt 3 = Tr A, Mt 4 = Tr D, Mt 5 = Tr G, Mt 6 = Tr H, Mt 7 = Tr C, Mt 8 = Tr B). A first indication of orthology might be that the clover BACs map to positions in the clover genome that are equivalent to the positions in the *Medicago* genome of their proposed orthologous *Medicago* sequence. All of the clover BACs satisfied this requirement insofar as they all map to the appropriate clover chromosome. Insufficient detail was available from the

publication of George *et al.* (2006) to extend the comparison beyond this level, and we await publication of the full study of this group to enable us to carry out a more precise analysis of the relative positions of the clover and *Medicago* sequences.

Using *M. truncatula* as a 'nodal' species for comparison, a growing number of studies have begun to reveal extensive synteny between the members of the legume family. In 2003, Yan *et al.* reported genome-wide conserved microsynteny between the genomes of *M. truncatula* and soybean, with 54% of 50 soybean contig groups showing conserved microsynteny to *M. truncatula* (Yan *et al.*, 2003). In 2004, the same group analysed eight homologous BAC contig groups in detail by comparative physical mapping and cross-hybridization and found six of eight genome regions exhibited conserved synteny, including three that were extensively conserved (Yan *et al.*, 2004). In the same year, Choi *et al.* carried out a preliminary analysis of the soybean *rgh1* region and concluded that 63% of the genes were conserved and collinear between soybean and *M. truncatula* (Choi *et al.*, 2004b). During the same study, the extent of conserved microsynteny between *M. truncatula* and *Lotus japonicus* was assessed with the analysis of ten clone-pairs with broadly genetic positions in the two genomes. Of the 91 and 84 genes identified in *M. truncatula* and *L. japonicus*, respectively, 72 genes (82%) were conserved homologs, with the majority present in conserved order and orientation (Choi *et al.*, 2004c). Due to the genome sequence availability of *M. truncatula* and *L. japonicus*, a more in depth analysis of similarity between these two species has been undertaken (Cannon *et al.*, 2006a). This study showed the existence of 10 large-scale synteny blocks, which account for 67% of the *M. truncatula* genome and 64% of *L. japonicus*. Across synteny blocks, "synteny quality" (calculated as twice the number of homologue matches divided by the total number of genes in both species, excluding transposable elements and collapsing tandem duplications) between the two species averaged 54% (Cannon *et al.*, 2006).

In our study, extensive conserved synteny was observed between all of the proposed clover and *Medicago* orthologues examined. Within the window of comparison defined by the clover BACs, a total of 93 and 96 genes were identified in white clover and *M. truncatula*, respectively, with 39 conserved homologues. There was complete conservation of gene order between the homologues in both species, and,

with only one exception (gene 11 in 27K12), all homologues had the same transcriptional orientation. Using the same method as Cannon *et al.* (2006), synteny quality between white clover and *Medicago* was 42% over all of the sequences (ranging from 33% to 60% in individual comparisons). However, the figure of 39 conserved homologues between the two species is based only on regions of nucleotide similarity corresponding to predicted genes. Amongst the five BACs, there are also seven instances of apparent intergenic similarity, which we hypothesise may represent orthologous genes not predicted by FGENESH (the order of these regions was also completely conserved between the species). If these are taken into account, synteny quality increases to 52%. This is doubtless an underestimate, for several reasons. Firstly, the clover BACs were sequenced only to a sixfold level of coverage, and during the comparison process, a number of smaller contigs (>6 Kb) were excluded from analysis. Thus sequencing coverage of the clover BACs was by no means complete and it is probable that some genes (with possible *Medicago* orthologues) have not been revealed. Secondly, the window of comparison between the two species was, by necessity, restricted according to the extent of the clover BAC sequence. Minor local rearrangements in either genome, which would be evident in the comparison of larger synteny blocks (as in the case of the *Medicago* – *Lotus* comparison) may not be evident in our data. While it might be tempting to compare the Cannon *et al.* study to our one, and question, (given the closer phylogenetic relationship of clover and *Medicago*, relative to *Lotus*), why the synteny quality between *Lotus* and *Medicago* is higher than that between clover and *Medicago*, the difference in scale and coverage between the two studies would render such a comparison of questionable value.

One of the main features of the white clover genome that we have not addressed during this study is its allotetraploid status. While it is likely that each of the five genomic intervals represented by the clover BACs are orthologus to the *Medicago* interval to which they have been compared, it is not known which of the two homoeologous genomes of white clover any of these BACs originate from. In evolutionary terms, genome expansion through allopolyploidy is frequently accompanied by a process of gene loss and subsequent functional divergence in the gene complement of the homoeologous genomes. Because white clover is a recognisable allotetraploid, the homoeologous genomes (as evidenced by genetic

mapping studies, such as that of Barrett *et al.,* 2004) are certainly largely co-linear on a macrosyntenic scale, and are likely to exhibit extensive conserved microsynteny. However, deviations in the homosequential nature of homoeologous chromosomes in allopolyploids have been observed in other species, such as wheat (Gale *et al.,* 1993), and it is highly likely that such deviations exist in white clover. It would be interesting to extend the current analysis to the examination of BACs that represent the 'alternative' homoeologue of the intervals of the white clover genome represented by the BACs analysed in this study. Initial reports also suggest that the white clover genome contains numerous paralogous duplications, complicating comparisons with closely related species such as *M. truncatula* (Cogan *et al.,* 2006; George *et al.,* 2006). As previously discussed, genetic mapping of SSRs derived from the BACs analysed in this study suggested that they did not possess any well preserved paralogues in the clover genome.

In our opinion, this study demonstrates the excellent potential utility of the genespace sequence of *Medicago truncatula* to advance white clover genomics and genetics. The work in this chapter has demonstrated that the BES strategy outlined in the previous chapter does indeed have potential to identify clover BAC and their potential orthologous regions in the *Medicago* genome, and that a scaled up version of the approach has the potential to form the basis of a powerful tool for gene discovery in white clover. In addition, a combination of high throughput BAC-end sequencing and targeted sequencing of individual BACs may yield valuable insights into the evolutionary history of white clover.

5.0 Conclusions

5.1 Conclusions

The large genome sizes of many important crop plants render map-based cloning experiments especially difficult. Thus, it is attractive to advance experimental strategies for gene isolation in species with large genomes by exploiting comparative mapping with a related species, which is characterised by a small genome. The Webster dictionary states that "translation implies the rendering from one language into another", while 'genomics' is generally considered the use of high throughput methods to study genomes both form and function. Therefore, translational genomics entails going from the language of genomics to that of breeding, or in other words, the direct application of genomic resources to make plant breeding programs easier and more efficient. Based on these ideas, the purpose of the work presented in this thesis was to assess several key factors required for the development of a truly useful platform for translational genomics between the model legume *M. truncatula* and the forage legume species, white clover (*Trifolium repens* L.).

Some studies have been performed to elucidate the extent of genome structure conservation between *M. truncatula* and crop legumes for map-based cloning of agronomically important genes, especially those required for nodulation in crop legumes, using *M. truncatula* as a surrogate genome. One example was the successful cloning of the pea orthologue of the *SYM2* gene characterised first in *M. truncatula* (Limpens *et al.*, 2003). Pea *SYM2* is a putative nodfactor entry receptor involved in the rhizobial infection process (Geurts *et al.*, 1997). Map-based cloning of *SYM2* in pea was difficult due to its large genome and the lack of efficient transformation methods in pea. However, the pea *SYM2* region is highly syntenic with *M. truncatula* (Gualtieri *et al.*, 2002). The strongly linked markers flanking the *SYM2* in pea were used to identify *M. truncatula* BACs, and a physical contig (approximately 300 kb) covering the *SYM2* orthologous region of *M. truncatula* was sequenced to identify candidate genes. Using the RNA interference reverse genetic tool, Limpens et al. (2003) showed that two LysM-domain receptor kinases were specifically involved in infection thread formation, and, therefore, are probable orthologs of the *SYM2* in pea.

The above study illustrates the power of translational genomics approaches in the legumes, for the isolation of agronomically important genes from species possessing

genomes that are large and complex. However, it also illustrates some of the barriers to the adoption of this strategy as a general tool for the use of model species for gene isolation in related crop species. Firstly, to adopt the approach described, on a gene-by-gene basis, without any a priori knowledge of the likelihood of the success of the strategy for a particular gene-target is very undesirable. It would be useful, at the outset of such an experiment, to have a very well developed idea of the similarity of the entire target genome to that of the model genome, both on a macrosyntenic and microsyntenic scale. This would both facilitate the ease with which such a translational positional cloning strategy could be undertaken, while identifying potential problems with factors such as distinguishing between orthologues and paralogues – a very real problem in plant genomes which tend to have an evolutionary history containing multiple rounds of whole genome duplication. Secondly, while the approach described for pea works well for genes that are expressed in a qualitative fashion, it would be much more challenging to employ this strategy on a regular basis to attempt to identify candidate genes underlying quantitative traits in the target agricultural species.

White clover could be said to epitomize many of the potential problems mentioned above - it is an allotetraploid species with a large genome of ca. 960 Mb. It is a self incompatible outbreeder, with a long breeding cycle (10-20 years), low heritabilities for many important traits and is affected by environmental parameters that make difficult selection regimes and reduce genetic gains. Thus, the genome of white clover is characterised by extensive whole genome duplication, and the vast majority of the traits which breeders and geneticists would like to dissect are quantitative in nature. All of this argues for the development of a strong, integrated set of genome-based tools that would allow successful translational genomics strategies in white clover (using *Medicago* as the nodal species).

Even though a number of studies have explored the comparative structural genomics of *M. truncatula* and other legumes, up until recently, comparisons between *Medicago* and white clover have been lacking. Recently, the situation has improved. For instance, Cogan *et al.,* (2006) demonstrated the relatively high levels of conservation of gene content between clover and *Medicago* based on in silico EST comparisons and using a genetic mapping approach, incorporating markers from

these ESTs; George *et al.,* (2006) have established broad homology between the chromosomes of white clover and those of *Medicago truncatula*. Both should allow the easier transferability of *M. truncatula* information onto the genome white clover, however while these studies have provided insights into the conservation of gene content and macrosynteny between clover and *Medicago*, no one has yet explored levels of genome wide microsynteny between these species.

The central premise of the study described in this thesis was to develop some of the tools that would be required for an extensive exploration of the extent of genome-wide microsynteny between clover and *Medicago*, and assess one possible strategy through which such a resource could be exploited. To this end, a bacterial artificial chromosome (BAC) library was constructed, spanning a three equivalent coverage of the genome of white clover. Each BAC represented (in our case) 85,000 nucleotides of clover sequence. One of the possible ways in which this BAC library could be exploited for the assessment of genome-wide macrosynteny is to identify what we refer to as comparative-tile-BACs, which represent probable orthologous sequence intervals between clover (BACs) and *Medicago* (genome sequence). Our pilot BAC-end sequencing study demonstrated that in silico comparison of large numbers of sequenced clover BAC-ends to the *Medicago* genespace sequence would result in many matches. However, we felt that it was unlikely that all of these represented orthologues, and so a second "layer" of information was used in order to identify the clover BACs most likely to represent orthologues of particular *Medicago* sequences, based on the requirement for both ends of a BAC to have matches in the *Medicago* genome, at a distance compatible with the length of a BAC. This very conservative approach discovered relatively few putative comparative tile BACs, but, assuming a linear increase in the results on scaling up to a larger study, we believe that the BAC-end sequencing approach is a useful tool to anchor a large amount of sequence of white clover onto the *Medicago* genome. Further analysis of sequence analysis of five white clover BAC clones showed large conserved segments between a model and a crop plant, and allowed us to demonstrate the probable orthologous nature of the clover BACs and comparable *Medicago* sequence, but we have pointed out that our study does not take into account the allotetraploid status of white clover.

During the course of the project we also developed a genetic linkage map in a cross involving the genotype in which the BAC library was constructed. The main role for this map was to provide a resource that would allow us to add a "macrosyntenic context" to some of the microsyntenic comparisons that we made during the course of the study. This contextualisation generally took the form of mapping SSRs derived from clover BAC-ends. However, since the resulting F_1 population segregated for several morphological characteristics of interest to breeders, these were scored over the entire population and a QTL mapping exercise was undertaken. While the majority of effects were reasonably small and multiple-season and environment replication was limited, we feel that in the longer term, mapping in multiple populations will help identify the loci which are consistently involved in these traits.

One thing that is illustrated quite well by the QTL mapping study in this thesis is the fact that, even with good translational genomics resources, it will be difficult to find the genes which underlie QTLs for truly polygenic traits in clover. However, quantitative genetic variation for morphological traits such as leaf dimensions and plant height, and for reproductive traits such as flowering date and seedpod weight, has been detected in *M. truncatula* (Bonnin *et al.*, 1996, 1997). The genetic control of reproductive traits such as components of seed yield was also analysed in alfalfa (Bolanos-Aguilar *et al.*, 2002). Because of the level of relatedness between white clover and these species, it is possible that the genetic control of some or all of these traits is similar in terms of genetic architecture amongst these species. Given the inbreeding nature and small genome size of *M. truncatula*, and the fact that a considerable array of functional genomics tools also exist for that species, it is much more likely that the genes underlying quantitative variation for such traits will be identified in *Medicago*. A well-developed, integrated translational genomics platform linking the *Medicago* and clover genomes might be invaluable in future initiatives to "translate" the dissection of the genetic architecture of quantitative traits in *Medicago* to white clover. Given the results during the course of this thesis, we feel confident in stating that combining the high throughput BAC-end sequencing strategy we outline in this thesis with tools such as a comprehensive comparative genetic map of white clover and *Medicago* will provide the powerful integrated platform for translational genomics that will allow the dissection of quantitative traits in white clover in the future.

Bibliography

Aagaard, J.E., Krutovskii, K.V., and Strauss, S.H. (1998). RAPDs and allozymes exhibit similar levels of diversity and differentiation among populations and races of Douglas-fir. *Heredity* **81**, 69-78.

Ahn, S., and Tanksley, S.D. (1993). Comparative Linkage Maps of the Rice and Maize Genomes. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 7980-7984.

Allard, R.W. (1956). The analysis of genetic-environmental interactions by means of diallel crosses. *Genetics* **41**, 305-318.

Alvarez, J., Guli, C.L., Yu, X.H., and Smyth, D.R. (1992). Terminal-Flower - a Gene Affecting Inflorescence Development in Arabidopsis-Thaliana. *Plant Journal* **2**, 103-116.

Andersen, J.R., and Lubberstedt, T. (2003). Functional markers in plants. *Trends in Plant Science* **8**, 554-560.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815.

Araki, T., and Komeda, Y. (1993). Analysis of the Role of the Late-Flowering Locus, Gl, in the Flowering of Arabidopsis-Thaliana. *Plant Journal* **3**, 231-239.

Australian Government (2004). The biology and ecology of white clover (Trifolium repens L.) in Australia. *Department of Health and Ageing*.

Axelsson, T., Shavorskaya, O., and Lagercrantz, U. (2001). Multiple flowering time QTLs within several Brassica species could be the result of duplicated copies of one ancestral gene. *Genome* **44**, 856-864.

Bardacki, F. (2001). Random amplified polymorphic DNA (RAPD) markers. *Turk J Biol* **25**, 185-196.

Barrett, B., Griffiths, A., Schreiber, M., Ellison, N., Mercer, C., Bouton, J., Ong, B., Forster, J., Sawbridge, T., Spangenberg, G., Bryan, G., and Woodfield, D. (2004). A microsatellite map of white clover. *Theoretical and Applied Genetics* **109**, 596-608.

Barrett, B., Mercer, C., and Woodfield, D. (2005a). Genetic mapping of a root-knot nematode resistance locus in Trifolium. *Euphytica* **143**, 85-92.

Barrett, B.A., Baird, I.J., and Woodfield, D.R. (2005b). A QTL analysis of white clover seed production. *Crop Science* **45**, 1844-1850.

Barth, S., Gonzales, M., Febrer, M., and Connolly, V. (2004). Molecular genetic diversity within and among Irish ecotypes of perennial ryegrass and white clover collected from old pastures. *In* "Genetic variation for plant breeding" (H. G. P. R. J. Vollmann, ed.), pp. 162.

Becker, J., and Heun, M. (1995). Barley Microsatellites - Allele Variation and Mapping. *Plant Molecular Biology* **27**, 835-845.

Bell, C.J., and Ecker, J.R. (1994). Assignment of 30 Microsatellite Loci to the Linkage Map of Arabidopsis. *Genomics* **19**, 137-144.

Bennetzen, J.L., and Freeling, M. (1997). The unified grass genome: Synergy in synteny. *Genome Research* **7**, 301-306.

Bennetzen, J.L., and Ma, J.X. (2003). The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. *Current Opinion in Plant Biology* **6**, 128-133.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580.

Blanc, G., and Wolfe, K.H. (2004). Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *Plant Cell* **16**, 1667-1678.

Bolanos-Aguilar, E.-D., Huyghe, C., Ecalle, C., Hacquet, J., and Julier, B. (2002). Effect of Cultivar and Environment on Seed Yield in Alfalfa. *Crop Sci* **42**, 45-50.

Bonierbale, M.W., Plaisted, R.L., and Tanksley, S.D. (1988). Rflp Maps Based on a Common Set of Clones Reveal Modes of Chromosomal Evolution in Potato and Tomato. *Genetics* **120**, 1095-1103.

Bonnin, I., Prosperi, J.M., and Olivieri, I. (1996). Genetic markers and quantitative genetic variation in *Medicago* truncatula (Leguminosae): A comparative analysis of population structure. *Genetics* **143**, 1795-1805.

Bonnin, I., Prosperi, J.M., and Olivieri, I. (1997). Comparison of quantitative genetic parameters between two natural populations of a selfing plant species, *Medicago* truncatula Gaertn. *Theoretical and Applied Genetics* **94**, 641-651.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a Genetic-Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *American Journal of Human Genetics* **32**, 314-331.

Boutin, S.R., Young, N.D., Olson, T.C., Yu, Z.H., Shoemaker, R.C., and Vallejos, C.E. (1995). Genome Conservation among 3 Legume Genera Detected with DNA Markers. *Genome* **38**, 928-937.

Bouton, J.H. (1996). New uses of alfalfa and other "old" forage legumes. *In* "Progress in new crops" (J. Janick, ed.), pp. 251-259. ASHS Press, Alexandria, VA.

Bowley, S.R. (1997). Breeding methods for forage legumes. *In* "Biotechnology and the improvement of forage legumes" (B. D. a. B. McKersie, D.C.W., ed.), pp. 25-42. CAB International.

Broten, K. (2000). Microsatellites. In "Agbiotech Infosource", Issue 55, Ag-West Biotech Inc, Canada.

Broun, P., and Tanksley, S.D. (1996). Characterization and genetic mapping of simple repeat sequences in the tomato genome. *Molecular & General Genetics* **250**, 39-49.

Brouwer, D.J., and Osborn, T.C. (1999). A molecular marker linkage map of tetraploid alfalfa (*Medicago* sativa L.). *Theoretical and Applied Genetics* **99**, 1194-1200.

Brudno, M., Chapman, M., Gottgens, B., Batzoglou, S., and Morgenstern, B. (2003a). Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics* **4**, 66.

Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I., and Batzoglou, S. (2003b). Glocal Alignment: Finding Rearrangements During Alignment. *Bioinformatics* **19S1**, i54-i62.

Burr, B. (2002). Mapping and Sequencing the Rice Genome. *Plant Cell* **14**, 521-523.

Burr, B., Burr, F.A., Thompson, K.H., Albertson, M.C., and Stuber, C.W. (1988). Gene Mapping with Recombinant Inbreds in Maize. *Genetics* **118**, 519-526.

Cannon, S.B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., Wang, X., Mudge, J., Vasdewani, J., Schiex, T., Spannagl, M., Monaghan, E., Nicholson, C., Humphray, S.J., Schoof, H., Mayer, K.F.X., Rogers, J., Quetier, F., Oldroyd, G.E., Debelle, F., Cook, D.R., Retzel, E.F., Roe, B.A., Town, C.D., Tabata, S., Van de Peer, Y., and Young, N.D. (2006). Legume genome evolution viewed through the *Medicago* truncatula and Lotus japonicus genomes. *PNAS* **103**, 14959-14964.

Caradus, J.R., and Chapman, D.F. (1991). Variability of Stolon Characteristics and Response to Shading in 2 Cultivars of White Clover (Trifolium-Repens L). *New Zealand Journal of Agricultural Research* **34**, 239-247.

Caradus, J.R., and Chapman, D.F. (1996). Selection for and heritability of stolon characteristics in two cultivars of white clover. *Crop Science* **36**, 900-904.

Caradus, J.R., and Mackay, A.C. (1991). Performance of White Clover Cultivars and Breeding Lines in a Mixed Species Sward .2. Plant Characters Contributing to Differences in Clover Proportion in Swards. *New Zealand Journal of Agricultural Research* **34**, 155-160.

Chang, C., Bowman, J.L., Dejohn, A.W., Lander, E.S., and Meyerowitz, E.M. (1988). Restriction Fragment Length Polymorphism Linkage Map for

Arabidopsis-Thaliana. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 6856-6860.

Chen, M., SanMiguel, P., deOliveira, A.C., Woo, S.S., Zhang, H., Wing, R.A., and Bennetzen, J.L. (1997). Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 3431-3435.

Chen, M.S., Presting, G., Barbazuk, W.B., Goicoechea, J.L., Blackmon, B., Fang, F.C., Kim, H., Frisch, D., Yu, Y.S., Sun, S.H., Higingbottom, S., Phimphilai, J., Phimphilai, D., Thurmond, S., Gaudette, B., Li, P., Liu, J.D., Hatfield, J., Main, D., Farrar, K., Henderson, C., Barnett, L., Costa, R., Williams, B., Walser, S., Atkins, M., Hall, C., Budiman, M.A., Tomkins, J.P., Luo, M.Z., Bancroft, I., Salse, J., Regad, F., Mohapatra, T., Singh, N.K., Tyagi, A.K., Soderlund, C., Dean, R.A., and Wing, R.A. (2002). An integrated physical and genetic map of the rice genome. *Plant Cell* **14**, 537-545.

Chen, M.S., SanMiguel, P., and Bennetzen, J.L. (1998). Sequence organization and conservation in sh2/a1-homologous regions of sorghum and rice. *Genetics* **148**, 435-443.

Choi, H.K., Kim, D.J., Uhm, T., Limpsens, E., Lim, H., Mun, J.H., Kalo, P., Penmetsa, R.V., Seres, A., Kulikova, O., Roe, B.A., Bisseling, T., Kiss, G.B., and Cook, D.R. (2004a). A Sequence-based genetic map of *Medicago* truncatula and comparison of marker colinearity with *Medicago* sativa. *Genetics* **166**, 1463-1502.

Choi, H.K., Mun, J.H., Kim, D.J., Zhu, H.Y., Baek, J.M., Mudge, J., Roe, B., Ellis, N., Doyle, J., Kiss, G.B., Young, N.D., and Cook, D.R. (2004b). Estimating genome conservation between crop and model legume species. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 15289-15294.

Choi, S.D., Creelman, R.A., Mullet, J.E., and Wing, R.A. (1995). Construction and characterization of a bacterial artificial chromosome library of Arabidopsis thaliana. *Weeds World* **2**, 17-20.

Chou, H.H., and Holmes, M.H. (2001). DNA sequence quality trimming and vector removal. *Bioinformatics* **17**, 1093-1104.

Chung, S.M., and Staub, J.E. (2003). The development and evaluation of consensus chlroroplast primer pairs that possess highly variable sequence regions in a diverse array of plant taxa. *Theoretical and Applied Genetics* **107**, 757-767.

Clarke, J.H., and Dean, C. (1994). Mapping Fri, a Locus Controlling Flowering Time and Vernalization Response in Arabidopsis-Thaliana. *Molecular & General Genetics* **242**, 81-89.

Clarke, L., and Carbon, J. (1976). A colony bank containing Col El hybrid plasmids representative of the entire E.coli genome. *Cell* **9**, 91-99.

Cobos, M.J., Fernandez, M., Rubio, J., Kharrat, M., Moreno, M.T., Gil, J., and Millan, T. (2005). A linkage map of chickpea (Cicer arietinum L.) based on populations from Kabuli x Desi crosses: location of genes for resistance to fusarium wilt race 0. *Theoretical and Applied Genetics* **110**, 1347-1353.

Cogan, N.O.I., Abberton, M.T., Smith, K.F., Kearney, G., Marshall, A.H., Williams, A., Michaelson-Yeates, T.P.T., Bowen, C., Jones, E.S., Vecchies, A.C., and Forster, J.W. (2006). Individual and multi-environment combined analyses identify QTLs for morphogenetic and reproductive development traits in white clover (Trifolium repens L.). *Theoretical and Applied Genetics* **112**, 1401-1415.

Cogan, N.O.I., Drayton, M.C., Ponting, R.C., Vecchies, A.C., Bannan, N.R., Sawbridge, T.I., Smith, K.F., Spangenberg, G.C., and Forster, J.W. (2007). Validation of in silico-predicted genic SNPs in white clover (Trifolium repens L.), an outbreeding allopolyploid species. *Molecular Genetics and Genomics* **277**, 413-425.

Collins, A., Lonjou, C., and Morton, N.E. (1999). Genetic epidemiology of single-nucleotide polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 15173-15177.

Connolly, V. (2001). "Breeding improved varieties of white clover." Teagasc Crops Research Centre, Carlow.

Cook, D.R., Van den Bosch, K.A., de Bruijn, F.J., and Huguet, T. (1997). Model legumes get the nod. *The Plant Cell*, 275-281.

Crush, J.R. (1987). Nitrogen fixation. *In* "White clover" (M. A. Baker and W. M. Williams, eds.), pp. 185-201. C.A.B. International.

Danesh, D., Penuela, S., Mudge, J., Denny, R.L., Nordstrom, H., Martinez, J.P., and Young, N.D. (1998). A bacterial artificial chromosome library for soybean and identification of clones near a major cyst nematode resistance gene. *Theoretical and Applied Genetics* **96**, 196-202.

d'Erfurth, I., Cosson, V., Mondy, S., Brocard, L., Kondorosi, A., and Ratet, P. (2006). The low level of activity of Arabidopsis thaliana Tag1 transposon correlates with the absence of two minor transcripts in *Medicago* truncatula. *Molecular Breeding* **17**, 317-328.

Devos, K., and Gale, M. (2000). Plant Comparative Genetics after 10 Years. *Science* **282**, 656.

Devos, K.M., and Gale, M.D. (1997). Comparative genetics in the grasses. *Plant Molecular Biology* **35**, 3-15.

Ding, Y., Johnson, M.D., Colayco, R., Chen, Y.J., Melnyk, J., Schmitt, H., and Shizuya, H. (1999). Contig assembly of bacterial artificial chromosome

clones through multiplexed fluorescence-labeled fingerprinting. *Genomics* **56**, 237-246.

Dograr, N., and Akkaya, M.S. (2001). Optimization of PCR amplification of wheat simple sequence repeat DNA markers. *Turk J Biol* **25**, 153-158.

Doniskeller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., Lander, E.S., Botstein, D., Akots, G., Rediker, K.S., Gravius, T., Brown, V.A., Rising, M.B., Parker, C., Powers, J.A., Watt, D.E., Kauffman, E.R., Bricker, A., Phipps, P., Mullerkahle, H., Fulton, T.R., Ng, S., Schumm, J.W., Braman, J.C., Knowlton, R.G., Barker, D.F., Crooks, S.M., Lincoln, S.E., Daly, M.J., and Abrahamson, J. (1987). A Genetic-Linkage Map of the Human Genome. *Cell* **51**, 319-337.

Edwards, A., Voss, H., Rice, P., Civitello, A., Stegemann, J., Schwager, C., Zimmermann, J., Erfle, H., Caskey, C.T., and Ansorge, W. (1990). Automated DNA Sequencing of the Human Hprt Locus. *Genomics* **6**, 593-608.

Ellison, N.W., Liston, A., Steiner, J.J., Williams, W.M., and Taylor, N.L. (2006). Molecular phylogenetics of the clover genus (Trifolium - Leguminosae). *Molecular Phylogenetics and Evolution* **39**, 688-705.

Endre, G., Kereszt, A., Kevei, Z., Mihacea, S., Kalo, P., and Kiss, G.B. (2002). A receptor kinase gene regulating symbiotic nodule development. *Nature* **417**, 962-966.

Eujayl, I., Baum, M., Powell, W., Erskine, W., and Pehu, E. (1998). A genetic linkage map of lentil (Lens sp.) based on RAPD and AFLP markers using recombinant inbred lines. *Theoretical and Applied Genetics* **97**, 83-89.

Evan, L.T. (1993). In "Crop evolution, adaptation and yield" Cambridge, University Press, UK.

Febrer, M., Cheung, F., Town, C.D., Cannon, S.B., Young, N.D., Abberton, M., Jenkins, G., and Milbourne, D. (2007). Construction, characterization and preliminary BAC-end sequencing analysis of a bacterial artificial chromosome library of white clover (Trifolium repens L.). *Genome* **50**, 412-421.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Research* **32**, W273-9.

Freeling, M. (2001). Grasses as a single genetic system. Reassessment 2001. *Plant Physiology* **125**, 1191-1197.

Friesen, L. (1998). Arabidopsis: Model plant in biotech research. In "Agbiotech Infosourse, Issue 40, Ag-West Biotech Inc, Canada.

Frugoli, J., and Harris, J. (2001). *Medicago* truncatula on the Move! *The Plant Cell*, 458-462.

Gale, M.D., Atkinson, M.D., Chinoy, C.N., Harcourt, R.L., Jia, J., and al., e. (1993). Genetic maps of hexaploid wheat. *In* "8th International Genetic Symposium" (Z. S. Li and Z. Y. Xin, eds.), Beijing.

Gale, M.D., and Devos, K.M. (1998). Plant comparative genetics after 10 years. *Science* **282**, 656-659.

Gao, M.Q., Li, G.Y., McCombie, W.R., and Quiros, C.F. (2005). Comparative analysis of a transposon-rich Brassica oleracea BAC clone with its corresponding sequence in A-thaliana. *Theoretical and Applied Genetics* **111**, 949-955.

Gaut, B.S., and Doebley, J.F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 6809-6814.

George, J., Cogan, N.O.I., Smith, K.F., Spangenberg, G.C., and Forster, J.W. (2006). Genetic map integration and comparative genome organisation of white clover (Trifolium repens L.) with model legumes. *In* "Plant & Animal Genome XIV", pp. 214, San Diego, CA.

Geurts, R., Heidstra, R., Hadri, A.E., Downie, J.A., Franssen, H., vanKammen, A., and Bisseling, T. (1997). Sym2 of pea is involved in a nodulation factor-perception mechanism that controls the infection process in the epidermis. *Plant Physiology* **115**, 351-359.

Green, E.D. (2001). Strategies for the systematic sequencing of complex genomes. *Nature Reviews Genetics* **2**, 573-583.

Gualtieri, G., Kulikova, O., Limpens, E., Kim, D.J., Cook, D.R., Bisseling, T., and Geurts, R. (2002). Microsynteny between pea and *Medicago* truncatula in the SYM2 region. *Plant Molecular Biology* **50**, 225-235.

Gupta, P.K., Roy, J.K., and Prasad, M. (2001). Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science* **80**, 524-535.

Halliday, K.J., Koornneef, M., and Whitelam, G.C. (1994). Phytochrome B and at Least One Other Phytochrome Mediate the Accelerated Flowering Response of Arabidopsis-Thaliana L to Low Red/Far-Red Ratio. *Plant Physiology* **104**, 1311-1315.

Hamada, H., Petrino, M.G., and Kakunaga, T. (1982). A Novel Repeated Element with Z-DNA-Forming Potential is Widely Found in Evolutionarily Diverse Eukaryotic Genomes. *PNAS* **79**, 6465-6469.

Han, O.K., Kaga, A., Isemura, T., Wang, X.W., Tomooka, N., and Vaughan, D.A. (2005). A genetic linkage map for azuki bean Vigna angularis (Willd.) Ohwi & Ohashi. *Theoretical and Applied Genetics* **111**, 1278-1287.

Harborne, J.B. (1994). Phytochemistry of the Leguminosae. *In* "Phytochemical dictionary of the Leguminosae" (F. A. e. a. Bisby, ed.). Chapman & Hall, London.

Harrison, M.J. (2000). Molecular genetics of model legumes. *Trends in Plant Science* **5**, 414-415.

Hauge, B.M., Hanley, S.M., Cartinhour, S., Cherry, J.M., Goodman, H.M., Koornneef, M., Stam, P., Chang, C., Kempin, S., Medrano, L., and Meyerowitz, E.M. (1993). An Integrated Genetic Rflp Map of the Arabidopsis-Thaliana Genome. *Plant Journal* **3**, 745-754.

Hayashi, M., Miyahara, A., Sato, S., Kato, T., Yoshikawa, M., Taketa, M., Pedrosa, A., Onda, R., Imaizumi-Anraku, H., Bachmair, A., Sandal, N., Stougaard, J., Murooka, Y., Tabata, S., Kawasaki, S., Kawaguchi, M., and Harada, K. (2001). Construction of a genetic linkage map of the model legume Lotus japonicus using an intraspecific F-2 population. *DNA Research* **8**, 301-310.

Hecht, V., Foucher, F., Ferrandiz, C., Macknight, R., Navarro, C., Morin, J., Vardy, M.E., Ellis, N., Beltran, J.P., Rameau, C., and Weller, J.L. (2005). Conservation of Arabidopsis Flowering Genes in Model Legumes. *Plant Physiol.* **137**, 1420-1434.

Herrmann, D., Boller, B., Studer, B., Widmer, F., and Kolliker, R. (2006). QTL analysis of seed yield components in red clover (Trifolium pratense L.). *Theoretical and Applied Genetics* **112**, 536-545.

Holliday, R. (1989). "The economics of grass/white clover for dairy cows."

Huala, E., and Sussex, I.M. (1992). Leafy Interacts with Floral Homeotic Genes to Regulate Arabidopsis Floral Development. *Plant Cell* **4**, 901-913.

Huguet, T., and Prosperi, J.M. (1996). "*Medicago* truncatula: a legume model-plant." Ressource from the International Centre for Advanced Mediterrenean Agronomic Studies.

Hymowitz, T. (1990). Grain legumes. *In* "Advances in new crops" (J. a. S. Janick, J.E., ed.), pp. 54-57. Timber press, Portland, OR.

Isobe, S., Klimenko, I., Ivashuta, S., Gau, M., and Kozlov, N.N. (2003). First RFLP linkage map of red clover (Trifolium pratense L.) based on cDNA probes and its transferability to other red clover germplasm. *Theoretical and Applied Genetics* **108**, 105-112.

Jahufer, M.Z.Z., Cooper, M., Ayres, J.F., and Bray, R.A. (2002). Identification of research to improve the efficiency of breeding strategies for white clover in

Australia - a review. *Australian Journal of Agricultural Research* **53**, 239-257.

Jones, C. (2005). Molecular genetic dissection of complex traits in white clover. Master Degree, University of Wales Aberystwyth, Aberystwyth.

Jones, C.J., Edwards, K.J., Castaglione, S., Winfield, M.O., Sala, F., vandeWiel, C., Bredemeijer, G., Vosman, B., Matthes, M., Daly, A., Brettschneider, R., Bettini, P., Buiatti, M., Maestri, E., Malcevschi, A., Marmiroli, N., Aert, R., Volckaert, G., Rueda, J., Linacero, R., Vazquez, A., and Karp, A. (1997). Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. *Molecular Breeding* **3**, 381-390.

Jones, E.S., Hughes, L.J., Drayton, M.C., Abberton, M.T., Michaelson-Yeates, T.P.T., Bowen, C., and Forster, J.W. (2003). An SSR and AFLP molecular marker-based genetic map of white clover (Trifolium repens L.). *Plant Science* **165**, 531-539.

Julier, B., S., F., Barre, P., Cardinet, G., Santoni, S., Huguet, T., and Huyghe, C. (2003). Construction of two genetic linkage maps in cultivated tetraploid alfalfa (*Medicago* sativa) using microsatellite and AFLP markers. *BMC Plant Biology* **3**.

Kalo, P., Endre, G., Zimanyi, L., Csanadi, G., and Kiss, G.B. (2000). Construction of an improved linkage map of diploid alfalfa (*Medicago* sativa). *Theoretical and Applied Genetics* **100**, 641-657.

Kao, C.H., Zeng, Z.B., and Teasdale, R.D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203-1216.

Kashkush, K., Feldman, M., and Levy, A.A. (2002). Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* **160**, 1651-1659.

Klein, R.R., Morishige, D.T., Klein, P.E., Dong, J., and Mullet, J.E. (1998). High throughput BAC DNA isolation for physical map construction of sorghum (Sorghum bicolor). *Plant Molecular Biology Reporter* **16**, 351-364.

Knight, W.E. (1953). Interrelationships of Some Morphological and Physiological Characteristics of Ladino Clover. *Agronomy Journal* **45**, 197-199.

Kowalski, S.P., Lan, T.H., Feldmann, K.A., and Paterson, A.H. (1994). Comparative Mapping of Arabidopsis-Thaliana and Brassica-Oleracea Chromosomes Reveals Islands of Conserved Organization. *Genetics* **138**, 499-510.

Ku, H.-M., Vision, T., Liu, J., and Tanksley, S.D. (2000). Comparing sequenced segments of the tomato and Arabidopsis genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *PNAS* **97**, 9121-9126.

Kulikova, O., Gualtieri, G., Geurts, R., Kim, D.J., Cook, D., Huguet, T., de Jong, J.H., Fransz, P.F., and Bisseling, T. (2001). Integration of the FISH pachytene and genetic maps of *Medicago* truncatula. *Plant Journal* **27**, 49-58.

Kumar, L.S. (1999). DNA markers in plant improvement: An overview. *Biotechnology Advances* **17**, 143-182.

Kurata, N., Moore, G., Nagamura, Y., Foote, T., Yano, M., Minobe, Y., and Gale, M. (1994). Conservation of Genome Structure between Rice and Wheat. *Bio-Technology* **12**, 276-278.

Lagercrantz, U., Ellegren, H., and Andersson, L. (1993). The Abundance of Various Polymorphic Microsatellite Motifs Differs between Plants and Vertebrates. *Nucleic Acids Research* **21**, 1111-1115.

Lagercrantz, U., and Lydiate, D.J. (1996). Comparative genome mapping in Brassica. *Genetics* **144**, 1903-1910.

Lagercrantz, U., Putterill, J., Coupland, G., and Lydiate, D. (1996). Comparative mapping in Arabidopsis and Brassica, fine scale genome collinearity and congruence of genes controlling flowering time. *Plant Journal* **9**, 13-20.

Lan, T.-H., DelMonte, T.A., Reischmann, K.P., Hyman, J., Kowalski, S.P., McFerson, J., Kresovich, S., and Paterson, A.H. (2000). An EST-enriched Comparative Map of Brassica oleracea and Arabidopsis thaliana. *Genome Res.* **10**, 776-788.

Lander, E.S., and Botstein, D. (1989). Mapping Mendelian Factors Underlying Quantitative Traits Using Rflp Linkage Maps. *Genetics* **121**, 185-199.

Lee, J.M., Grant, D., Vallejos, C.E., and Shoemaker, R.C. (2001). Genome organization in dicots. II. Arabidopsis as a 'bridging species' to resolve genome evolution events among legumes. *Theoretical and Applied Genetics* **103**, 765-773.

Li, W., and Gill, B.S. (2002). The Colinearity of the Sh2/A1 Orthologous Region in Rice, Sorghum and Maize Is Interrupted and Accompanied by Genome Expansion in the Triticeae. *Genetics* **160**, 1153-1162.

Limpens, E., Franken, C., Smit, P., Willemse, J., Bisseling, T., and Geurts, R. (2003). LysM domain receptor kinases regulating rhizobial Nod factor-induced infection. *Science* **302**, 630-633.

Limpens, E., Ramos, J., Franken, C., Raz, V., Compaan, B., Franssen, H., Bisseling, T., and Geurts, R. (2004). RNA interference in Agrobacterium rhizogenes-transformed roots of Arabidopsis and *Medicago* truncatula. *Journal of Experimental Botany* **55**, 983-992.

Lotti, C., Salvi, S., Pasqualone, A., Tuberosa, R., and Blanco, A. (2000). Integration of AFLP markers into an RFLP-based map of durum wheat. *Plant Breeding* **119**, 393-401.

Maliepaard, C., Alston, F.H., van Arkel, G., Brown, L.M., Chevreau, E., Dunemann, F., Evans, K.M., Gardiner, S., Guilford, P., van Heusden, A.W., Janse, J., Laurens, F., Lynn, J.R., Manganaris, A.G., den Nijs, A.P.M., Periam, N., Rikkerink, E., Roche, P., Ryder, C., Sansavini, S., Schmidt, H., Tartarini, S., Verhaegh, J.J., Vrielink-van Ginkel, M., and King, G.J. (1998). Aligning male and female linkage maps of apple (Malus pumila Mill.) using multi-allelic markers. *Theoretical and Applied Genetics* **97**, 60-73.

Marshall, E. (1999). A high-stakes gamble on genome sequencing. *Science* **284**, 1906-1909.

Meinke, D.W., Cherry, J.M., Rounsley, S.D., and Koornneef, M. (1998). Arabidopsis thaliana: a model plant for genome analysis. *Science* **282**, 662-682.

Meksem, K., Zobrist, K., Ruben, E., Hyten, D., Quanzhou, T., Zhang, H.B., and Lightfoot, D.A. (2000). Two large-insert soybean genomic libraries constructed in a binary vector: applications in chromosome walking and genome wide physical mapping. *Theoretical and Applied Genetics* **101**, 747-755.

Men, A.E., Meksem, K., Kassem, M.A., Lohar, D., Stiller, J., Lightfoot, D., and Gresshoff, P.M. (2001). A bacterial artificial chromosome library of Lotus japonicus constructed in an Agrobacterium tumefaciens-transformable vector. *Molecular Plant-Microbe Interactions* **14**, 422-425.

Menancio-Hautea, D., Fatokun, C.A., Kumar, L., Danesh, D., and Young, N.D. (1993). Comparative Genome Analysis of Mungbean (Vigna-Radiata L-Wilczek) and Cowpea (V-Unguiculata L Walpers) Using Rflp Mapping Data. *Theoretical and Applied Genetics* **86**, 797-810.

Menendez, C.M., Hall, A.E., and Gepts, P. (1997). A genetic linkage map of cowpea (Vigna unguiculata) developed from a cross between two inbred, domesticated lines. *Theoretical and Applied Genetics* **95**, 1210-1217.

Michelmore, R.W., Paran, I., and Kesseli, R.V. (1991). Identification of Markers Linked to Disease-Resistance Genes by Bulked Segregant Analysis - a Rapid Method to Detect Markers in Specific Genomic Regions by Using Segregating Populations. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 9828-9832.

Miyao, A., Zhong, H.S., Monna, L., Yano, M., Yamamoto, K., Havukkala, I., Minobe, Y., and Saski, T. (1996). Characterization and genetic mapping of simple sequence repeats in the rice genome. *DNA Research* **3**, 233-238.

Moore, G., Devos, K.M., Wang, Z., and Gale, M.D. (1995). Cereal Genome Evolution - Grasses, Line up and Form a Circle. *Current Biology* **5**, 737-739.

Morgante, M., Rafalski, A., Biddle, P., Tingey, S., and Olivieri, A.M. (1994). Genetic-Mapping and Variability of 7 Soybean Simple Sequence Repeat Loci. *Genome* **37**, 763-769.

Mozo, T., Dewar, K., Dunn, P., Ecker, J.R., Fischer, S., Kloska, S., Lehrach, H., Marra, M., Martienssen, R., Meier-Ewert, S., and Altmann, T. (1999). A complete BAC-based physical map of the Arabidopsis thaliana genome. *Nature Genetics* **22**, 271-275.

Mueller, U.G., and LaReesa Wolfenbarger, L. (1999). AFLP genotyping and fingerprinting. *Tree* **14**, 389-394.

Mylona, P., Pawlowski, K., and Bisseling, T. (1995). Symbiotic Nitrogen-Fixation. *Plant Cell* **7**, 869-885.

Nagamura, Y., Horiuchi, I., Sugioka, K., Watanabe, K., Antonio, B.A., Akimoto, M., Matoba, N., Numa, H., Honda, S., Shimuzu, Y., and Higo, K. (2003). Rice-BLAST: a comprehensive homology searcg for rice specific sequences. *Genome Informatics* **14**, 533-534.

Nakamura, Y., Asamizu, E., and Kaneto, T. (2002). A legume Lotus japonicus genome association. *Genome Informatics* **13**, 539-540.

Newbury, H.J. (2003). "Plant molecular breeding," School of Sciences, University of Birmingham, UK.

Osborn, T.C., Kole, C., Parkin, I.A.P., Sharpe, A.G., Kuiper, M., Lydiate, D.J., and Trick, M. (1997). Comparison of flowering time genes in Brassica rapa, B-napus and Arabidopsis thaliana. *Genetics* **146**, 1123-1129.

Ouedraogo, J.T., Gowda, B.S., Jean, M., Close, T.J., Ehlers, J.D., Hall, A.E., Gillaspie, A.G., Roberts, P.A., Ismail, A.M., Bruening, G., Gepts, P., Timko, M.P., and Belzile, F.J. (2002). An improved genetic linkage map for cowpea (Vigna unguiculata L.) Combining AFLP, RFLP, RAPD, biochemical markers, and biological resistance traits. *Genome* **45**, 175-188.

Paterson, A.H., Bowers, J.E., Burow, M.D., Draye, X., Elsik, C.G., Jiang, C.X., Katsar, C.S., Lan, T.H., Lin, Y.R., Ming, R.G., and Wright, R.J. (2000). Comparative genomics of plant chromosomes. *Plant Cell* **12**, 1523-1539.

Paterson, A.H., Bowers, J.E., and Chapman, B.A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9903-9908.

Pedrosa, A., Sandal, N., Stougaard, J., Schweizer, D., and Bachmair, A. (2002). Chromosomal Map of the Model Legume Lotus japonicus. *Genetics* **161**, 1661-1672.

Polhill, R.M. (1981). Sophoreae. *In* "Advances in Legume Systematics. Part I" (R. M. P. P. H. Raven, ed.), pp. 213-230. Royal Botanic Gardens.

Powell, W., Machray, G.C., and Provan, J. (1996). Polymorphism revealed by simple sequence repeats. *Trends in Plant Science* **1**, 215-222.

Powell, W., Thomas, W., Thompson, D.M., Swanston, J.S., and Waugh, R. (1992). Association Between rDNA Alleles and Quantitative Traits in Doubled Haploid Populations of Barley. *Genetics* **130**, 187-194.

Putterill, J., Robson, F., Lee, K., Simon, R., and Coupland, G. (1995). The Constans Gene of Arabidopsis Promotes Flowering and Encodes a Protein Showing Similarities to Zinc-Finger Transcription Factors. *Cell* **80**, 847-857.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000). The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research* **28**, 141-145.

Rabinowicz, P.D., and Bennetzen, J.L. (2006). The maize genome as a model for efficient sequence analysis of large plant genomes. *Current Opinion in Plant Biology* **9**, 149-156.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* **5**, 94-100.

Rafalski, J.A., and Tingey, S.V. (1993). Genetic Diagnostics in Plant-Breeding - Rapds, Microsatellites and Machines. *Trends in Genetics* **9**, 275-280.

Redei, G.P. (1962). Supervital Mutants of Arabidopsis. *Genetics* **47**, 443-460.

Reiter, R., Williams, J., Feldmann, K., Rafalski, J., Tingey, S., and Scolnik, P. (1992). Global and Local Genome Mapping in Arabidopsis thaliana by Using Recombinant Inbred Lines and Random Amplified Polymorphic DNAs. *PNAS* **89**, 1477-1481.

Ritter, E., Gebhardt, C., and Salamini, F. (1990). Estimation of Recombination Frequencies and Construction of RFLP Linkage Maps in Plants From Crosses Between Heterozygous Parents. *Genetics* **125**, 645-654.

Sandal, N., Krusell, L., Radutoiu, S., Olbryt, M., Pedrosa, A., Stracke, S., Sato, S., Kato, T., Tabata, S., Parniske, M., Bachmair, A., Ketelsen, T., and Stougaard, J. (2002). A genetic linkage map of the model legume Lotus japonicus and strategies for fast mapping of new loci. *Genetics* **161**, 1673-1683.

Sasaki, T., Matsumoto, T., Antonio, B.A., and Nagamura, Y. (2005). From Mapping to Sequencing, Post-sequencing and Beyond. *Plant Cell Physiol.* **46**, 3-13.

Sato, S., Isobe, S., Asamizu, E., Ohmido, N., Kataoka, R., Nakamura, Y., Kaneko, T., Sakurai, N., Okumura, K., Klimenko, I., Sasamoto, S., Wada, T., Watanabe, A., Kohara, M., Fujishiro, T., and Tabata, S. (2005).

Comprehensive Structural Analysis of the Genome of Red Clover (Trifolium pratense L.). *DNA Res* **12**, 301-364.

Sato, S., Kaneko, T., Nakamura, Y., Asamizu, E., Kato, T., and Tabata, S. (2001). Structural analysis of a Lotus japonicus genome. I. Sequence features and mapping of fifty-six TAC clones which cover the 5.4 Mb regions of the genome. *DNA Research* **8**, 311-318.

Sato, S., and Tabata, S. (2006). Lotus japonicus as a platform for legume research. *Current Opinion in Plant Biology* **9**, 128-132.

Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* **8**, 552-560.

Scholte, M., d'Erfurth, I., Rippa, S., Mondy, S., Cosson, V., Durand, P., Breda, C., Trinh, H., Rodriguez-Llorente, I., Kondorosi, E., Schultze, M., Kondorosi, A., and Ratet, P. (2002). T-DNA tagging in the model legume *Medicago* truncatula allows efficient gene discovery. *Molecular Breeding* **10**, 203-215.

Senior, M.L., and Heun, M. (1993). Mapping Maize Microsatellites and Polymerase Chain-Reaction Confirmation of the Targeted Repeats Using a Ct Primer. *Genome* **36**, 884-889.

Sheaffer, C., Mathison, R., Martin, N., Rabas, D., and Ford, H. (2003). "Forage Legumes: Second Edition--Clovers, Birdsfoot Trefoil, Cicer Milkvetch, Crownvetch, Sainfoin and Alfalfa."

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992). Cloning and Stable Maintenance of 300-Kilobase-Pair Fragments of Human DNA in Escherichia-Coli Using an F-Factor-Based Vector. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 8794-8797.

Shoemaker, R.C., Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J.P., Kochert, G., and Boerma, H.R. (1996). Genome Duplication in Soybean (Glycine subgenus soja). *Genetics* **144**, 329-338.

Song, Q.J., Marek, L.F., Shoemaker, R.C., Lark, K.G., Concibido, V.C., Delannay, X., Specht, J.E., and Cregan, P.B. (2004). A new integrated genetic linkage map of the soybean. *Theoretical and Applied Genetics* **109**, 122-128.

Stracke, S., Kistner, C., Yoshida, S., Mulder, L., Sato, S., Kaneko, T., Tabata, S., Sandal, N., Stougaard, J., Szczyglowski, K., and Parniske, M. (2002). A plant receptor-like kinase required for both bacterial and fungal symbiosis. *Nature* **417**, 959-962.

Tadege, M., Ratet, P., and Mysore, K.S. (2005). Insertional mutagenesis: a Swiss army knife for functional genomics of *Medicago* truncatula. *Trends in Plant Science* **10**, 229-235.

Tanksley, S.D. (1993). Mapping polygenes. *Annual Review of Genetics* **27**, 205-233.

Tanksley, S.D., Bernatzky, R., Lapitan, N.L., and Prince, J.P. (1988). Conservation of Gene Repertoire but not Gene Order in Pepper and Tomato. *PNAS* **85**, 6419-6423.

Tanksley, S.D., Ganal, M.W., Prince, J.P., Devicente, M.C., Bonierbale, M.W., Broun, P., Fulton, T.M., Giovannoni, J.J., Grandillo, S., Martin, G.B., Messeguer, R., Miller, J.C., Miller, L., Paterson, A.H., Pineda, O., Roder, M.S., Wing, R.A., Wu, W., and Young, N.D. (1992). High-Density Molecular Linkage Maps of the Tomato and Potato Genomes. *Genetics* **132**, 1141-1160.

Tanksley, S.D., Young, N.D., Paterson, A.H., and Bonierbale, M.W. (1989). Rflp Mapping in Plant-Breeding - New Tools for an Old Science. *Bio-Technology* **7**, 257-264.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S., and Rafalski, A. (2000). The Complete Sequence of 340 kb of DNA around the Rice Adh1-Adh2 Region Reveals Interrupted Colinearity with Maize Chromosome 4. *Plant Cell* **12**, 381-392.

Tautz, D., and Renz, M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucl. Acids Res.* **12**, 4127-4138.

Thiel, T., Michalek, W., Varshney, R.K., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics* **106**, 411-422.

Thoday, J.M. (1961). Location of polygenes. *Nature* **191**, 368–370.

Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L., and Avramova, Z. (1999). Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 7409-7414.

Tomkins, J.P., Yu, Y., Miller-Smith, H., Frisch, D.A., Woo, S.S., and Wing, R.A. (1999). A bacterial artificial chromsome library for sugarcane. *Theoretical and Applied Genetics* **99**, 419-424.

Town, C.D., Cheung, F., Maiti, R., Crabtree, J., Haas, B.J., Wortman, J.R., Hine, E.E., Althoff, R., Arbogast, T.S., Tallon, L.J., Vigouroux, M., Trick, M., and Bancroft, I. (2006). Comparative Genomics of Brassica oleracea and Arabidopsis thaliana Reveal Gene Loss, Fragmentation, and Dispersal after Polyploidy. *Plant Cell* **18**, 1348-1359.

Van Ooijen, J.W., Boer, M.P., Jansen, R.C., and Maliepaard, C. (2002). MapQTL® 4.0, Software for the calculation of QTL positions on genetic maps. *Plant Research International, Wageningen, the Netherlands*.

Van Ooijen, J.W., and Voorrips, R.E. (2001). JoinMap® 3.0, Software for the calculation of genetic linkage maps. *Plant Research International, Wageningen, The Netherlands*.

VandenBosch, K.A., and Stacey, G. (2003). Summaries of Legume Genomics Projects from around the Globe. Community Resources for Crops and Models. *Plant Physiol.* **131**, 840-865.

Vandepoele, K., Saeys, Y., Simillion, C., Raes, J., and Van de Peer, Y. (2002). The Automatic Detection of Homologous Regions (ADHoRe) and Its Application to Microcolinearity Between Arabidopsis and Rice. *Genome Res.* **12**, 1792-1801.

Vandepoele, K., Simillion, C., and Van de Peer, Y. (2003). Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* **15**, 2192-2202.

Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O., and Hunkapiller, M. (1998). Shotgun sequencing of the human genome. *Science* **280**, 1540-1542.

Vignal, A., Milan, D., SanCristobal, M., and Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution* **34**, 275-305.

Vision, T.J., Brown, D.G., and Tanksley, S.D. (2000). The origins of genomic duplications in Arabidopsis. *Science* **290**, 2114-2117.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Vandelee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., and Zabeau, M. (1995). Aflp - a New Technique for DNA-Fingerprinting. *Nucleic Acids Research* **23**, 4407-4414.

Weeden, N.F., Muehlbauer, F.J., and Ladizinsky, G. (1992). Extensive Conservation of Linkage Relationships between Pea and Lentil Genetic Maps. *Journal of Heredity* **83**, 123-129.

Weigel, D. (1995). The genetics of flower development:from floral induction to ovule morphogenesis. *Annual Review of Genetics* **29**, 19-39.

White, S., and Doebley, J. (1998). Of genes and genomes and the origin of maize. *Trends in Genetics* **14**, 327-332.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A., and Tingey, S.V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucl. Acids Res.* **18**, 6531-6535.

Williams, W.M. (1987). Genetics and Bredding. *In* "White clover" (M. J. Baker and W. M. Williams, eds.), pp. 343-416. C.A.B International.

Wu, K.S., and Tanksley, S.D. (1993). Abundance, Polymorphism and Genetic-Mapping of Microsatellites in Rice. *Molecular & General Genetics* **241**, 225-235.

Wyman, A., and White, R. (1982). A Highly Polymorphic Locus in Human DNA. *Cytogenetics and Cell Genetics* **32**, 329-329.

Yan, H.H., Mudge, J., Kim, D.J., Larsen, D., Shoemaker, R.C., Cook, D.R., and Young, N.D. (2003). Estimates of conserved microsynteny among the genomes of Glycine max, *Medicago* truncatula and Arabidopsis thaliana. *Theoretical and Applied Genetics* **106**, 1256-1265.

Yan, H.H., Mudge, J., Kim, D.J., Shoemaker, R.C., Cook, D.R., and Young, N.D. (2004). Comparative physical mapping reveals features of microsynteny between Glycine max, *Medicago* truncatula, and Arabidopsis thaliana. *Genome* **47**, 141-155.

Young, N.D. (1996). QTL mapping and quantitative disease resistance in plants. *Annual Review of Phytopathology* **34**, 479-501.

Young, N.D., Cannon, S.B., Sato, S., Kim, D., Cook, D.R., Town, C.D., Roe, B.A., and Tabata, S. (2005). Sequencing the genespaces of *Medicago* truncatula and Lotus japonicus. *Plant Physiology* **137**, 1174-1181.

Young, N.D., Mudge, J., and Ellis, T.H.N. (2003). Legume genomes: more than peas in a pod. *Current Opinion in Plant Biology* **6**, 199-204.

Young, N.D., and Shoemaker, R.C. (2006). Genome studies and molecular genetics Part 1 Model legumes exploring the structure, function and evolution of legume genomes - Editorial overview. *Current Opinion in Plant Biology* **9**, 95-98.

Young, N.E. (1989). Review of legume systems for beef and sheep. In "The attractions of Forage Legumes" (British Grassland Society, Ed), Hurley, UK.

Yu, J., Wang, J., Lin, W., Li, S.G., Li, H., Zhou, J., Ni, P.X., Dong, W., Hu, S.N., Zeng, C.Q., Zhang, J.G., Zhang, Y., Li, R.Q., Xu, Z.Y., Li, S.T., Li, X.R., Zheng, H.K., Cong, L.J., Lin, L., Yin, J.N., Geng, J.N., Li, G.Y., Shi, J.P., Liu, J., Lv, H., Li, J., Deng, Y.J., Ran, L.H., Shi, X.L., Wang, X.Y., Wu, Q.F., Li, C.F., Ren, X.Y., Wang, J.Q., Wang, X.L., Li, D.W., Liu, D.Y., Zhang, X.W., Ji, Z.D., Zhao, W.M., Sun, Y.Q., Zhang, Z.P., Bao, J.Y., Han, Y.J., Dong, L.L., Ji, J., Chen, P., Wu, S.M., Liu, J.S., Xiao, Y., Bu, D.B., Tan, J.L., Yang, L., Ye, C., Zhang, J.F., Xu, J.Y., Zhou, Y., Yu, Y.P., Zhang, B., Zhuang, S.L., Wei, H.B., Liu, B., Lei, M., Yu, H., Li, Y.Z., Xu, H., Wei, S.L., He, X.M., Fang, L.J., Zhang, Z.J., Zhang, Y.Z., Huang, X.G., Su, Z.X., Tong, W., Li, J.H., Tong, Z.Z., Li, S.L., Ye, J., Wang, L.S., Fang, L., Lei, T.T., Chen, C., Chen, H., Xu, Z., Li, H.H., Huang, H.Y., Zhang, F., Xu, H.Y., Li, N., Zhao, C.F., Dong, L.J., Huang, Y.Q., Li, L., Xi, Y., Qi, Q.H., Li, W.J.,

Hu, W., Zhang, Y.L., Tian, X.J., Jiao, Y.Z., et al. (2005). The Genomes of Oryza sativa: A history of duplications. *Plos Biology* **3**, 266-281.

Zabeau, M., and Vos, P. (1993). Selective restriction fragment amplification: A general method for DNA fingerprinting. European Patent Office, publication 0 534 858 A1.

Zhang, Y., Sledge, M.K., and Bouton, J.H. (2007). Genome mapping of white clover (Trifolium repens L.) and comparative analysis within the Trifolieae using cross-species SSR markers. *Theoretical and Applied Genetics* **114**, 1367-1378.

Zhao, W., Wang, J., He, X., Huang, X., Jiao, Y., Dai, M., Wei, S., Fu, J., Chen, Y., Ren, X., Zhang, Y., Ni, P., Zhang, J., Li, S., Wang, J., Wong, G.K.S., Zhao, H., Yu, J., Yang, H., and Wang, J. (2004). BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Research* **32**, D377-D382.

Zhu, H., Choi, H.-K., Cook, D.R., and Shoemaker, R.C. (2005). Bridging Model and Crop Legumes through Comparative Genomics. *Plant Physiol.* **137**, 1189-1196.

Zhu, H.Y., Kim, D.J., Baek, J.M., Choi, H.K., Ellis, L.C., Kuester, H., McCombie, W.R., Peng, H.M., and Cook, D.R. (2003). Syntenic relationships between *Medicago* truncatula and Arabidopsis reveal extensive divergence of genome organization. *Plant Physiology* **131**, 1018-1026.

Appendices

Appendix A. List of white clover SSR from Barrett *et al*. (2004) used in the F_1(R3R4 x S1S4) mapping population. (Ats = genomic SSR; Prs = EST-SSR).

SSR ID	SSR Motif	Linkage group	Forward primer (5' - 3')	Reverse primer (5' - 3')
Ats029	ATG	A1, A2	GCTCCTTCCATTACGGTGTT	ATCGCTCTCGCTCTCCTTCT
Ats032	ATG	A1, A2	GGGCAGCCATTGTCATTC	TCGCAGTTATTCCGATCGATT
Ats054	CA	A1	GACACCGATTATGTGCAAGA	AATCACGACGAGCGACAACA
Prs461	TTC	A1	ACCTTCCGATATCCCAAACC	ATGGTGCGTTTGGAGATAGG
Prs426	TTG	A2	TTATTGATATCGGAAGCGACG	GTTCTCATGCGACGCTACAA
Ats131	CA	B1, B2	ATGATCCTCCGGACCGGATG	CGCACGGTGGACGTTGTAAG
Ats055	CA	B2, E1	CAATACAATCACCGCACCAG	TCTCTGCTTCGCGTCTTCTC
Ats121	ATG	B2	TCCACCGCTGTTGTGCAACC	TCTCTGCTTCGCGTCTTCTC
Ats067	CA	B2	TTACTGATGCGGTTAGTGTT	CCTCCTTGCTGTTGTTATCC
Ats075	CA	B1	ACTCGATCACCATGTGAGTC	TCATAACCGGCGTGGAAGAA
Ats205	CA	B1	ACACCGTCGTTATGAAGTAT	GACTTGTGACTACTGATAGA
Ats226	ATG	B1	ACACCGTCGTTATGAAGTAT	GACTTGTGACTACTGATAGA
Ats070	CA	C1, C2, G1	GTCATTGGTGATGGTGTTCT	TTTCGTCAGTGGCGGTGCTC
Prs264	ACC	C1	TCATCTTCTTCACCAGCACG	CGGTGAAACTGTTCCTGGTT
Prs433	TTC	C2	CCCAACCGCATAACACTTCT	AAAGATGAGGAGAAGATCAAGGG
Ats003	ATG	D1	CCTCAAGTCCACCACCTGTC	ACCACACCATTCTTGCTTCT
Prs599	TAA	D1	TTGAAGTTCCACAAAGATGTGC	AAACGACGTCACCAACCTTC
Prs427	CTT	D1	GTTCTTCAATCTCCACAATACGC	TAAAGGAAAAGGTGAGGATAGTGG
Prs612	ATG	D2	TTGAACTAGTCGTTGGATGGG	GAGAGGGTTTCAGGAACATACG
Ats113	ATG	D2	TGTGAGCTGGTGAATTGAGT	GGAGGTGATGATCTCTATCC
Prs247	AAG	D2	TCATCTTCATCAACAGTTTCCG	CTTCCCTTCTATCTCTCATGTTAACC
Ats058	CA	E1	CAATCAACTCTGCTAAGTGT	AAGGAGAGTATGTGAAGTAG
Prs408	TCT	E1	ACTAAGAAGGATCTGAATCTCTCTGC	TCACTGATCCAGTGGTAGATGG
Prs285	AAC	E2	TCCTCAAACGACCTCGTTCT	TCCTTCCAGGTCTTGCAGTT
Ats125	CTAT	E1	ATCTAAGGCTCCAAAGTATC	GCAACAATAGAAGCAGCAATCA
Prs256	TCTG	E1	CCGTTTTCGTTCTCGAAGAG	GAGTGGAGAGGAAGTCGTCG
Prs499	CCT	E2	TATGACTAGCCCCGTCATCC	AATTCCTTCGGGGAAGTTGT
Ats176	CA	F1, F2	ATCAGTTGGCGGTTCAGTAG	GTCAATGTGGCAGCTCTTGT
Ats123	CTAT	F1, F2	ACACAATTCAGCAGAGATTC	CAGTCGTTGGATCGTAGTAA
Prs006	CTT	G2	CCTGGAACTCTGAACTCGTACC	AGTGAATGGAGAAGAAGTGTGATG
Prs203	TGTT	G1	AACGATCCGATCTTGATTGC	ACTCATTTTTCCCTGCATGG
Prs510	TTC	G1	GCTGCTCTCCCTTTTTTCCT	GGAACATTTGCAGGGAGGTA
Prs305	AAC	G2	TCAAACTCGACGAAGAAGACG	GATGGAGATGGAGATTCGGA
Prs279	ATG	H1	ATGGCTAATGCAAAATTGGC	TAGAGACTGCAGATCATCATCTCC
Prs129	CCA	H2	CCGTGATTGCATATCCAGTG	CGCCACCCTTGTTAGTTGTT
Prs651	GAA	H1	GTAACGATTCTGTTGTTTACGTCG	CTCGTCGGTAAACGAATACTCC
Prs251	AGT	H2	TGGAGAAATTGGTGGTGTCA	TTCCACAACACCATCTCATCA
Ats066	CA	H1, H2	TATTCACCACACGCCTCTAC	ATGAGGAGAAGGCAGGAGAT
Ats186	CA	H1, H2	TGATGGCGCAATCAGGAATG	AACGCACACAGCCCTAGTTT

Appendix B. List of 95 PCR-based markers used in this study (from Choi *et al*. 2004a).

Name	Type	Function	Forward primer sequence (5' – 3')	Reverse primer sequence (5' – 3')
CysPr1	ESTe	Cysteine proteinase-like protein	GAGAATTCAAAGAAGAAATTAAGACAAAGA	GAAGAATTCATGGGGAGCAAAGT
PESR1	ESTe	Pectinesterase	CATCTGAACAAACCCATCTCCA	GCTGTTAATTCGGCGTTTGA
38K1L	BEST	Resistance gene analog	GGCTGGTATGAAAGAAGAGCAGAA	AGTTTAAATAAGGACCAGGATGTTCG
CDC16	ESTe	Cell division control protein 16	CCTCCCGCTTCACTTCACTTT	GGTAATGGTGGCCGAGGAATA
41O18L	BEST	Resistance gene analog	AGATATATCAGAAAAAACTAACCCAACCTT	AATACCCTTCCCTTTCCTTCCC
7H15L	BEST	Resistance gene analog	TTTCATCTTCTGCCTCATTTGGTC	GCCTATGGAGGTGAGGATTTGG
DK358L	BEST	Gm-RFLP-A363	GTTTGCGCCACTTAAGGTTATCTCATT	TGTCACCATGTGGCACATTCATT
DK238R	BEST	Gm-RFLP-A315	GCAATTTAAATGTAATCCATTGAACCA	TTATGCTTCTGATTCTAACTAACCCCA
DK407L	BEST	Gm-RFLP-A086	TTAATTTTATCAACCCACCATATTAGTCAA	CCAGTGCTGGAAAAGACAATCAATC
DK242R	BEST	Gm-RFLP-A947	CGTATGTTTAATCCGTTAGTCCGTCTT	CGTATGTTTAATCCGTTAGTCCGTCTT
DK353L	BEST	Gm-RFLP-A110	CCATGCCATGGAAGGGTGTTT	GCAAGAACCAGATACCCTTGACATTT
AW256557	ESTe	Resistance gene analog	GATATTTTCATTACTCAGCAACTTTTTCACAG	TGCTTCATCCCACTTATCATCAATACC
DK273L	BEST	Gm-RFLP-K300	TATGCCTGGTCTGTTCTTTCTTTACG	GCCCGTCCACCGCTTTTA
DK326R	BEST	Gm-RFLP-A064	CCAGCATGTAAACAATTGAAAGGCA	GTTGAACGGCTTAAATATCGCACTA
DK297L	BEST	Gm-RFLP-A656	GGGAAACACATGAGCGAAGGAGT	GCATAGCAAAACCACAATCTAACCA
DK321L	BEST	Gm-RFLP-A233	GAGCGAGCTCAGGATAGACTTTAGAA	TCCCACCTCCAATTTGTAGACGAT
DK296L	BEST	Gm-RFLP-K390	GAAAGGATGAGAAGCGGGGATAC	TCGTCGATGAAAAAGTACCAATAGAA
DK274L	BEST	Gm-RFLP-K390	TGCATAAGCTCAAAAATAAGTCAATCC	AGTAGATAAGCCCACATAAGCTCAAAATA
DK347L	BEST	Gm-RFLP-A063	AGATTTCATACCAGACGGAGGATAGTTC	TTTAGGTGATGGTGGCGTTGTTC
DK264L	BEST	Gm-RFLP-A688	GGGGAGTGTTGAGATATGCGTAAT	TGATCGAGGAACCAAAATAAAGAAA
DK501R	BEST	Gm-RFLP-A597	TATTTGGGATGGAAGCTATGTTGATTGG	TGCTTTAAAGGAGAAGGTAGATGATGAT
DK298R	BEST	Gm-RFLP-B139	ATAAGATAAGGGCCAACATAAGTAGAAAA	GAAACTTGAGAGTGAAGAAAGTGATAGAAC
NCAS	ESTi	Neuronal calcium sensor 1	TTCCCAAGCCCAAATCCTAAT	CATCACCAGGCCATCATCATAAGT
MPP	ESTi	Mitochondrial processing peptidase	TCCCCGAAACAATCCTCATCTG	GCAAATGTGTAGCCCCAAAAGTTA
DK009R	BEST	Ms-syntaxin-CG13	TAGCATCATCTTTCCCATACAA	GGGCAGGCAGCACCAGATA
DK006R	BEST	Pea-PTO-like kinase	GAACATAACCCCGAAGTGGAT	GAGTTTGGGAACAAAATTAGTATGAT
DK258L	BEST	Gm-RFLP-K007	GTATTCAGGGATTGAGTAAGAAAAAGGA	ACAAAATCCGTGGATGTATAAAAGTGTA
DK351L	BEST	Gm-RFLP-A110	TGCTTGGGCTTGAGCTTTTAGAA	CTGTTTGGGTATTAGTTTTTGTTGGG

ENOD8	ESTi	Enod 8	CCATGCCCATTCCTACTTTTCA	GTGGATTCCACGGACTTTACTTACT
RNAH	ESTi	ATP-dependent RNA helicases	GCTTCCACCAGCTGATACACG	TTAGCCCTAGCAAGAATGTCACTG
VBP1	ESTe	TGA-type basic leucine zipper protein	CTGGAGAGCAGACCCATTCAAT	GCGAAAGCCTCCAATCCAC
EST400	ESTi	Unknown	GGTGGCTGTCCCACTGATTATGT	AAATGCTTGTGTTATGCGGAGAG
TE019	ESTi	Unknown protein	GCATGTGACCGATGAGGAAACC	TTTTAGAATCAACAATGCAACCAGAAG
DNABP	ESTi	SAR DNA-binding protein	CCCTATGAGCTTGGGTTTGTCT	CTCATGGCATACGTGTTCAGC
MRS	ESTi	Methionyl-tRNA synthetase	GTCTGTGGTGGGATCATGGAGT	TTTTGACCGGTTCCAAGTAGAGTAG
MDH2	ESTi	Malate dehydrogenase	CTTCCATTTTCGATTCCTTTCATT	GCATGCCTCGACAACATCAGT
UNK7	ESTe	Putative protein	AAAAAGCAGCAAGAGAAATGTCAAT	GAGAATCTTTCTCCATCGTATCTTACTT
chit1	GS	Mt Chitinase I	GGTAAGGTAATGCTCTATCTTAATC	CTTACGGATGAAAGGTATTGTTTCC
SCP	ESTe	Serine carboxy-peptidase II	CACCAGCAGGAGAATCAAGGAAC	TCGATTCGTACCCAATTTGTTTTTCAATTGTT
ppPF	ESTe	Ppi-dependent phosphofructokinase sub	TCTCGCCACCAACAACAACTAC	AAAAATTGTTCATGAACACTCACTTGAAGCCA
DK020R	BEST	Mt TL4 probe	TCCATATTCAAGCCACCAATTCCCAT	ATACAAAATGTTACACTAAAACACGATA
GSb	GS	Glutamine synthetase	CTATGAGAGAAGATGGTGGCTATGAAGTCATCTTG	GGAGAGAACAATATTATTATTTGCTTACC
EST763	ESTi	Hypotetical protein	CACTCTAAAAAGGCCCAGAAGGTTTGACT	CTTATGACCAATAGTCTGTTCCACTC
EST158	ESTe	Vaculor sorting receptor-like protein	TTACAAACCACACCATAATTGCCAAATTG	TTGGTGACAACTGACACGAATGAAACTAC
RL13	ESTe	Ribosomal protein L13	GCCAAGCAGGTATCTATTCTTCATCT	CGGTACACAACATAACCCTAAAATCA
DENP	ESTe	Dentin phosphoryn [*Homo sapiens*]	AGAATTGGACTTCTTCTCACTCACG	CGGATGAAAAGCCTGAAGATAAGTC
11O9L	BEST	Resistance gene analog	CATGGCATCCAGATCCCACAT	TATAGACTTAGCCCTCAAAGTATTTCCC
DK277R	BEST	Gm-RFLP-A748	CTCAAATTCTCTAGTTTCAACATGGTATCA	GGGCTGTAGTATTTATACCTGAGTTAGTGAG
74O5R	BEST	Resistance gene analog	AGTTCATTACTTGATTAGCACACTTGTACA	AGTCTAGAATGGAACACGTTGTTTCG
EST758	ESTe	Hypotetical protein	TCACTTCCCCTAAATACGCTTCT	CTTAATTTCAGCTGCCATTTCAAC
19L13L	BEST	Resistance gene analog	TGTGTACAACAACAACAAGATAGAGGAACATT	CCATGGTTAAATGGAAGTAGTAACTGCTCC
24D15R	BEST	Resistance gene analog	AATAATTGACGAGCTACCAGCATATG	TGGATTTGAATGTGATCTTTTGATTAA
PRTS	ESTe	20S proteosome β-subunit	CATAGCTACTTGATCTGAAAACTTGACA	TGGTGAACTTCACTACCATTACAACC
75D1L	BEST	Resistance gene analog	AAGCTTCCAATGATAGATGATCGTAGA	TGTGGTATCATTAGTCCTTCTCATCAGG
ACCO	ESTi	1-ACC oxidase	GAAGATGGCGCAAAAAGAAAGT	GAAGATGGCGCAAAAAGAAAGT

DK293R	BEST	Gm-RFLP-A748	ACTTACAAGGTTAGCGTCATTCTCCATC	GCTATCCCACCTTAAAATTTCTTCACAA
TRPT	ESTe	Triosephosphate translocator	AACCACAATCTTTTCTCCCATCTT	AACATTCAAAGCCCACCAAGTT
18D24R	BEST	Resistance gene analog	CAATCCTGATCTACTTAACCAAAATAACAGC	GGATGAAAACAGAGAACCGTGAAACAC
40L12R	BEST	Resistance gene analog	ATGACATACTTCAAGAAATAAACCACCAG	ATCCAAATCCCATCTCCAACAGG
DK369R	BEST	Gm-RFLP-A450	GGAACGTGGAGTTGTTGATGGTATTAT	GATGTAAAAACCTTTACACTTGATTGATTG
MtEIL1	BEST	Ethylene sensitive	GACATGTATCGGATTCTCACGAGC	CACCTGCGAGGTATTCAAACTGTAA
6M23L	BEST	Resistance gene analog	GTTAGTTTACCACTTTTGAGTAGTGTAAGCA C	TTAATGTTAGAGATTGAAGGTGAGAGAAAC
33I23R	BEST	Resistance gene analog	CTCTTCATTTATGAGTGTACTTGTCTTTCC	ATGAATAGCCGTGTTTTGGTGG
DK045R	BEST	Mt-chitinase III	TGGCAATATCCACCAAATCAAA	CGAACCCACGACCACAAGG
DK332R	BEST	Gm-RFLP-A095	GGAAAATTATAAGCCAAACAACAGTAAAG	GATGATAACAATCGGGGAAAATAATG
5J9L	BEST	Resistance gene analog	TCCTTTGGGAAGAATGGTAGAGG	CTCTGAAGAAGTATTTTCCTTCCTTGAC
AW774053	ESTe	Resistance gene analog	CCGGGTAGTAGGGTCATCATTACA	CACAGCCATTTTATTATCTCCTCTCAAC
BE187590	ESTe	Resistance gene analog	ATGCGGATAGAAGGGCTGATGA	CAATTGTCCGGTCTGCTCTTCC
25A23L	BEST	Resistance gene analog	TTTTATTTCGCGTTGTTTATTTGATTC	ACGCGCAGCAGCCATCC
26G3L	BEST	Resistance gene analog	TTCTGACCAATCCGAAGAGCAGTGA	TGGGGTTAGATTTTAGTTACATGTTTGACAC A
KCoAT	ESTi	3-Ketoacyl-CoA thiolase	TGCTACTGCGGGTGGTAGATTTA	CTCCAGCACCATCACTCACCT
18L14L	BEST	Resistance gene analog	CGTAACATTCTCATTATCGCTGCTAT	AAGTAATCCGGTGATTGATTTTTCTCC
43I21L	BEST	Resistance gene analog	GCTTTTGTTTTAATGCATTTCTTAGTGTTTC	TGGAGCCTCATGTGTTTCAAACG
11N17R	BEST	Resistance gene analog	CTACTCCCTGCACCTAACCATTCACG	GCACAATTATTCATCTCTTCCCAAA
13B3R	BEST	Resistance gene analog	CTCGTTGTAAAAAAGCGTTACCAAACAGA	GTATTCATGTTACACAAAATAAACGTCATTG AG
AW736703	ESTi	Resistance gene analog	AGCAAGGTATTCAACTTCTTTTCATCTT	GCAAATATTTCTTACCAGTTAGTTTGTGC
42J16R	BEST	Resistance gene analog	ACCTCTTCTATAGAGATAACTTGTGTAGCAA G	GATAAACTGGCATTCCATGACTTTCA
zwilik	BEST	Zwille-like gene	ATTTGAGTGTACCCATTGAGAT	TTTTGAAGATTTATTTGTAGAGTA
3F15R	BEST	Resistance gene analog	ATGTCACGAAAATAAGCATACAAATCCTTC	GTGATGTTGCTTCCAGATGAATGTGG
AW257289	ESTi	Resistance gene analog	CTTCGGACCTTCAGCAAAACACAG	CGGGTGACAGATTATTTGGTGACATC
1N1R	BEST	Converted AFLP marker	CATATTGTTAGATTTGTGG	GTGAGCGTTAAGTTGGTAGAG
DK377R	BEST	Gm-RFLP-A487	AGCAGCTAGCACGTGTCCTTTGA	CCGACTAATTATCAATATTGTCAACTACATC T
PFK	ESTi	Ppi-dependent phosphofructokinase β-subunit	TCCCACTGCAAATCATGTCAAAAC	ACACAAGTGGATATTGATGGTTAGACTAC

AW256656	ESTi	Resistance gene analog	CCCAGACAACATTTCCTTACTATCGTCA	CCAAGTAGTAGGCAAAACCCAACAAATT
DSI	ESTi	Disulfide isomerase P5 precursor	CCAAGACATCTTTGGTTTCATCC	ACTGCAGAATCACTTGCCGAGTT
AW256637	ESTi	Resistance gene analog	TTCACCTAATTTCCATCTATACCATCCATGT	TATTTGTTAGCTTTAGTGATCGCTGCTACAC
PGKI	ESTi	Phosphoglycerate kinase	GATGACTGTATTGGCGAGGAAGT	GTTCGACACGGCTCCAACTA
RL3	ESTi	Ribosomal protein L3	GACACGGTTCTTTGGGATTTCTC	CCTGGCTTTTCGACTTCTCTGAC
TUP	ESTi	Translationally controlled tumor protein	GAATGGGATGCTATGGGAAGTG	TGGATCAGTGGCACCATCTTTAT
SAMS	ESTi	S-adenosyl-methionine synthase 2	CATAGCAAAGCGGGTTCAATCT	GTCAGCATCAAGACCAACATCATC
RLPO	ESTi	60S acidic ribosomal protein PO	TCTTGGCCCTTTAATTTCCTCTC	AACCTTGTATTTAGCAACTTCTTCACTG
RBBP	ESTi	WD-40 repeat protein MS14	CAAGAGGACGCAAACCTAAACC	CACAATTCGCAATCACCAAAGTAT
ACL	ESTi	ATP citrate lyase	AAGGTTAAGACCGTATTTATTCCAACA	AGTCCAAATTCGTCCCCACTG
NPAC	ESTi	Putative nascent polypeptide-associated	TGGCTCCAGGTCCAGTTATTGA	TCGGCTCTTCTTCTCGCTTCT
APX	ESTi	Ascorbate peroxidase	ATCTTCGCCATTTTCCTTCTTAG	GCTTTGCCAAACACATCCCTC

ESTi= exon-derived/intron spanning markers, ESTe= exon-derived markers, BEST= BAC end-sequence-tagged markers, GS= gene-sequence markers.

Appendix C. List of the SSRs identified from the BAC-end sequence analysis and their corresponding primer sequences.

Table C.1. List of the SSRs identified in the BAC-end sequences.

ID SSR	nr	SSR type	SSR	size	start	end
WCBE005TF	1	p2	(TA)26	52	383	434
WCBE010TF	1	p2	(TA)20	40	135	174
WCBE028TR	1	p2	(AT)21	42	126	167
WCBE040TF	1	p2	(AT)24	48	411	458
WCBE073TF	1	p3	(CTA)18	54	517	570
WCBE086TF	1	p2	(TA)24	48	727	774
WCBE101TR	1	p3	(TTA)12	36	6	41
WCBE113TF	1	p2	(GT)10	20	162	181
WCBE158TR	1	p3	(TCA)9	27	197	223
WCBE160TF	1	p2	(GA)14	28	462	489
WCBE181TF	1	c	(TTA)21tctgttcaaaaaa(ATT)13	115	5	119
WCBE195TR	1	p2	(TA)15	30	3	32
WCBE209TF	1	p2	(TA)11	22	719	740
WCBE224TR	1	p2	(TA)20	40	285	324
WCBE229TR	1	p3	(TTG)14	42	62	103
WCBE229TR	2	p4	(ATTA)5	20	295	314
WCBE304TF	1	p2	(TC)16	32	713	744
WCBE307TF	1	p2	(AT)20	40	583	622
WCBE325TF	1	c	(TC)18tttc(TA)10	60	540	599
WCBE336TF	1	p2	(AT)13	26	116	141
WCBE344TR	1	p2	(TA)29	58	30	87
WCBE364TR	1	p3	(CAT)10	30	620	649
WCBE381TR	1	p3	(TAT)8	24	143	166
WCBE387TR	1	p4	(TAAC)5	20	120	139
WCBE414TRB	1	p2	(TA)23	46	413	458
WCBE468TF	1	p3	(TTA)13	39	553	591
WCBE513TF	1	p2	(GT)14	28	509	536
WCBE517TRB	1	p2	(AT)22	44	672	715
WCBE555TF	1	p4	(TTCA)7	28	839	866
WCBE566TRB	1	p2	(TA)30	60	531	590
WCBE578TF	1	p4	(ACCA)6	24	356	379
WCBE579TF	1	p2	(TA)31	62	618	679
WCBE582TRB	1	p3	(CTT)7	21	307	327
WCBE602TF	1	p2	(AT)16	32	533	564
WCBE655TF	1	p1	(T)22	22	292	313
WCBE692TF	1	c	(AT)11(GT)15	52	317	368
WCBE693TF	1	p2	(TC)10	20	194	213
WCBE712TRB	1	p2	(TA)20	40	524	563
WCBE747TRB	1	p2	(AT)17	34	342	375
WCBE773TRB	1	p4	(ATCT)5	20	360	379
WCBE776TRB	1	p3	(TGT)10	30	535	564
WCBE786TF	1	p2	(AT)23	46	834	879
WCBE791TRB	1	p2	(GA)23	46	277	322

Table C.2. Sequences of the BAC-end sequence primer pairs that contained SSRs.

SSR ID	Forward primer (5' – 3')	Reverse primer (5' – 3')
WCBE005TF	TGCATCTTATCTACGCACTTTTG	CAAAATTAAAGAATCAAACCATGA
WCBE010TF	CCGGTGTAATTTGATCTCTCA	AAAAATCAAAGGTCATGATCTAAATG
WCBE028TR	TGAATTGTGAATGAATGAAAA	TCATTCTCTTTCAAAATTTCTCCA
WCBE040TF	CGCGAGAATCAACAGAAACC	CACTTGGTGAAGTGGTCCAT
WCBE073TF	TTTTGTGGCGGTTTCTATTG	TTGCCAATTGCCATCATGTA
WCBE086TF	TCTGCATTAGGCTTGGTGTG	AAGAAAGGGTCACCACAAAA
WCBE113TF	AGTGTTTAATAATTGGTTAAATTTTGG	TTGCATTTTCTAATCTACACAATTTCT
WCBE158TR	TTGCGCTTTACCGTTATTCC	CGATGACGACGATGGTTACA
WCBE160TF	GGAGTATTTCCCACCCCACT	TTCTACACCCGCGAGAGTTC
WCBE209TF	TTGGAGTTATAAATGACAAAATTGC	CTTGCAATTGATCCTACATACG
WCBE224TR	TTGCATACTGTGTTTTTCTTTTTGA	TTTTACGCATTGTGCTCGTC
WCBE229TR1	TTCACACCCTATACAATGGAGTAGT	CCCATTTTTCTCAAATCACAAAG
WCBE229TR2	TGAACAATAAAGCTTTTGGCTAAG	CAAATGAGTGAATGCCATGA
WCBE307TF	TTTGGATGAATTTATTTTGGGTTT	GCCCCTGGGTAACAAATTCT
WCBE325TF	AATCCATTAGCGGACTTCCA	CAGTGTTTAACATCTTGAAACCATTTA
WCBE336TF	AAATCCATTTTGCAAGATAACCA	ATCTCAAGCAAAGTAATTTTCAAGG
WCBE364TR	GCTCAACCTCCCAAACAACT	CCAACTTTGCAATTAACTCATCA
WCBE381TR	ACCAAGCAAGAAAGGCAAGA	TGAATAATTTTCTTCATTGTTGTTTAC
WCBE387TR	TCACCAAGGGGTAATGGAAC	CAAGTTTGAATTATGGTGGAGTG
WCBE468TF	TGATACTGTCGCGATACGTTATTT	GCAAGCATATTGGCTTGTTTG
WCBE513TF	TCTGACTCAACATTCTGTTTTCA	AACAAAAACAGGTATCGAACAGC
WCBE517TRB	GCGGCCAATATTGTTTGAAG	GTCGATTTCGAGCAGATTGA
WCBE566TRB	GGCGGAGGCTAGGGTAAAATC	AAAGCGCTATTGAGAAGTAT
WCBE578TF	GATGGTGGGAACAATTTTGG	ATCCGGAACGATTCGTGTAG
WCBE579TF	TCTTTCTTGCATTATGGCTCA	TTTCTGGAGCTCGGTACCAC
WCBE582TRB	GCAATTGGATTACCGACGAA	AAAAGCGCGTTCAACAATCT
WCBE602TF	ACCACGTGGGATGGAGAGTA	AACTTGCTAATGCACACCTTCA
WCBE655TF	CCATTGCGTATATGGAGCTG	TTTTCTTTAATTATTTCTCATGCGTTA
WCBE692TF	CACCAGCTGTCATTCCAAAA	GTGTGTGGAAAGGCAATCAA
WCBE712TRB	GAACAGGAGTTGGACGCAAT	TTGCCATCGATTTCTGCTAA
WCBE747TRB	TTTTGCTGTAAGAACGCATTG	GAGTTCTATATGAATTAATGGGAATCA
WCBE773TRB	GGGTTGGATTGGAGAATCTG	GCCATCAAATATCCACACACA
WCBE776TRB	CATCTTGGTTTTCCAGAATGG	GCTGGGTTGTTGCCTAACTC
WCBE791TRB	GAACTAAATACTGACATGGGCAAT	CAATTGAAAATTTGAAAGGGATCT

Appendix D. BLASTp results of the analogues genes in white clover and *M. truncatula.*

Table D.1. White clover clone 27B12

Genes	BLASTp result against *Arabidopsis thaliana* database	E value
1	Putative DNA binding protein	4 e-96
2	-	
3	-	
4	Putative retroelement pol polyprotein	5 e-143
5	-	
6	Centromere protein	4 e-141
7	-	
8	N-acetyltransferase	9 e-43
9	N-acetyltransferase	5 e-44
10	N-acetyltransferase	3 e-46
11	Structural constituent of ribosome (40S ribosomal protein S3a)	3 e-117
12	Unknown protein (NP_188893.1)	2 e-32
13	-	
14	Selenium binding like protein	2 e-129
15	-	
16	-	

Table D.2. White clover clone 27I09

Genes	BLASTp result against *Arabidopsis thaliana* database	E value
1	ATP binding/ kinase/ protein kinase	0.0
2	Double-stranded RNA binding	3 e-68
3	CAB77061.1	4 e-31
4	-	
5	NP_195531.1 (Far-red impaired response protein)	4 e-83
6	-	
7	-	
8	Pepsin A	6 e-37
9	Pepsin A	4 e-27
10	Catalytic	0.0
11	Nucleotide binding	0.0
12	NP_187146.2 (embryonic cell protein)	2 e-45

Table D.3. White clover clone 27K12

Genes	BLASTp result against *Arabidopsis thaliana* database	E value
1	SRG3 (Senescence related gene 3)	4 e-66
2	-	
3	Metal ion binding	1 e-159
4	Metal ion binding	6 e-25
5	Structural constituent of ribosome/ transcription regulator	2 e-59
6	-	
7	-	
8	Anaphase promoting complex/ cyclosome subunit	0.0
9	PBS kinase serine/ threonine-protein kinase	1 e-131
10	NP_566505.1	4 e-107
11	-	
12	-	
13	-	
14	-	
15	Tam-3 like tranposon protein hAT element transposase	4 e-20
16	Transcription factor/ zinc ion binding	7 e-21
17	Tam-3 like tranposon protein hAT element transposase	3 e-28
18	-	
19	-	
20	F5M15.6	0.0
21	-	
22	-	
23	-	

Table D.4. White clover clone 28F22

Genes	BLASTp result against *Arabidopsis thaliana* database	E value
1	F21J9.22	1 e-12
2	-	
3	-	
4	-	
5	-	
6	-	
7	Camodulin binding/ triacylglycerol lipase	0.0
8	-	
9	Structural constituent of ribosome	7 e-27
10	-	
11	-	
12	-	
13	-	
14	BAF01410.1	9 e-33
15	Transcription factor	0.0
16	-	
17	D-alanyl-D alanine endopeptidase	4 e-167
18	-	
19	-	
20	Carboxylic ester hydrolase	6 e-38
21	-	
22	-	
23	-	
24	-	

Table D.5. White clover clone 28G20

Genes	BLASTp result against *Arabidopsis thaliana* database	E value
1	BAB11018.1	5 e-69
2	ATP binding/ ATP dependent helicase/ DNA binding	0.0
3	Inorganic diphosphatase/ Mg^{2+} ion binding/ pyrophosphatase	7 e-91
4	Disulfide oxidoreductase/ oxidoreductase	0.0
5	-	
6	-	
7	-	
8	-	
9	-	
10	-	
11	ACO18907_17	4 e-21
12	-	
13	YAP169; gibberellin 20 oxidase	7 e-08
14	YAP169; gibberellin 20 oxidase	1 e-33
15	Transcription factor	4 e-46
16	Ankyrin-like protein	3 e-27

Table D.6. *Medicago truncatula* clone AC146852

Genes	BLASTp result against *Arabidopsis thaliana* database	E value
39	-	
38	-	
37	-	
36	Lactoglutathione lyase	2 e-53
35	Lactoglutathione lyase	8 e-35
34	Unknown protein (At3g22510)	7 e-22
33	-	
32	T25N20.6	7 e-90
31	Unknown protein (NP_566710.1)	1 e-31
30	-	5 e-11
29	-	
28	Unknown protein (NP_180537.1)	6 e-170
27	Unknown protein (NP_188893.1)	6 e-29
26	Structural constituent of ribosome (40S ribosomal protein S3a-1)	2 e-117
25	-	
24	N-acetyltransferase	9 e-45
23	Putative N-acetyltransferase	1 e-39
22	Putative N-acetyltransferase	1 e-43
21	-	
20	-	
19	N-acetyltransferase	1 e-40
18	Centromere protein	0.0
17	Putative DNA binding protein	2 e-114
16	-	
15	Unknown protein (NP_564808.1)	5 e-24
14	-	
13	ATPbinding/kinase/protein kinase	2 e-169
12	Selenium binding like protein	1 e-122
11	-	
10	Chain A X-Ray structure of gene product from A.th.	5 e-46
9	-	
8	-	
7	Chain A X-Ray structure of gene product from A.th.	8 e-43
6	-	
5	-	
4	-	
3	-	
2	-	
1	GTP binding	4 e-37

Table D.7. *Medicago truncatula* contig sequence MTCON74 (from 172,021 bp to 245,221 bp)

Genes	BLASTp result against *Arabidopsis thaliana* database	E value
1	-	
2	ATP binding/ kinase/ protein kinase	0.0
3	Double stranded RNA binding	1 e-60
4	Pepsin A	4 e-07
5	-	
6	Catalytic	0.0
7	-	
8	-	
9	Nucleotide binding	0.0
10	-	
11	Unknown protein (NP_187146.2)	1 e-44
12	Unknown protein (NP_187146.2)	6 e-13
13	-	
14	-	
15	GCN5-related N-acetyltransferase (GNAT) family protein	6 e-51
16	-	

Table D.8. *Medicago truncatula* clone AC133780

Genes	BLASTp result against *Arabidopsis thaliana* database	E value
31	NP_567588.1	3 e-156
30	NP_190088.1/ structural constituent of ribosome	9 e-42
29	-	
28	-	
27	-	
26	-	
25	IPS1 ubiquitin protein ligase	0.0
24	-	
23	-	
22	-	
21	-	
20	-	
19	-	
18	NP_196824.1 ABA responsive protein like	5 e-77
17	BAB08282.1	4 e-61
16	Zinc ion binding	6 e-48
15	NP_566505.1	9 e-121
14	DNA binding	3 e-15
13	-	
12	PBS1 kinase/ serine threonine protein kinase	1 e-129
11	Metal ion binding	2 e-37
10	-	
9	Metal ion binding	2 e-164
8	-	
7	Copia type polyprotein	0.0
6	-	
5	-	
4	-	
3	-	
2	-	
1	SRG3 (Senescence related gene 3)	1 e-45

Table D.9. *Medicago truncatula* contig sequence MTCON5806 (from 29,281 bp to 164,701 bp)

Genes	BLASTp result against *Arabidopsis thaliana* database	E value
1	Putative non LTR retroelement reverse transcriptase	0.0
2	Integral membrane family protein	3 e-149
3	-	
4	F21J9.11	1 e-17
5	-	
6	-	
7	-	
8	-	
9	-	
10	-	
11	-	
12	FRS5 (FAR1-related sequence 5)/ zinc ion binding	6 e-80
13	-	
14	Camodulin binding/ triacylglycerol lipase	2 e-177
15	-	
16	Structural constituent of ribosome	4 e-46
17	Transcription factor	0.0
18	-	
19	D-alanyl D-alanine endopeptidase	5 e-169
20	-	
21	-	
22	Carboxylic ester hydrolase	7 e-117
23	-	
24	-	
25	-	
26	Carboxylic ester hydrolase	1 e-83
27	AAM20148.1	0.0

Table D.10. *Medicago truncatula* BAC clone AC152349

Genes	BLASTp result against *Arabidopsis thaliana* database	E value
1	BAB11018.1	3 e-79
2	NP_196959.2	4 e-150
3	GPI-anchor transamidase	0.0
4	-	
5	BAB11018.1	2 e-60
6	-	
7	CAB72466.1	6 e-46
8	-	
9	-	
10	ATP binding/ ATP dependent helicase/ DNA binding	7 e-45
11	-	
12	ATP binding/ ATP dependent helicase/ DNA binding	0.0
13	-	
14	Inorganic diphosphatase/ Mg^{2+} ion binding/ pyrophosphatase	2 e-19
15	Disulfide oxidoreductase/ oxidoreductase	0.0
16	Disulfide oxidoreductase/ oxidoreductase	3 e-50
17	-	
18	-	
19	-	
20	-	
21	-	
22	YAP169; gibberellin 20 oxidase	3 e-89
23	-	
24	-	
25	-	
26	YAP169; gibberellin 20 oxidase	8 e-97
27	Transcription factor	1 e-49
28	-	
29	-	
30	-	
31	Ankyrin-like protein	2 e-153

Appendix E. Genome publication

Construction, characterization, and preliminary BAC-end sequencing analysis of a bacterial artificial chromosome library of white clover (*Trifolium repens* L.)

Melanie Febrer, Foo Cheung, Christopher D. Town, Steven B. Cannon, Nevin D. Young, Michael T. Abberton, Glyn Jenkins, and Dan Milbourne

Abstract: White clover (*Trifolium repens* L.) is a forage legume widely used in combination with grass in pastures because of its ability to fix nitrogen. We have constructed a bacterial artificial chromosome (BAC) library of an advanced breeding line of white clover. The library contains 37 248 clones with an average insert size of approximately 85 kb, representing an approximate 3-fold coverage of the white clover genome based on an estimated genome size of 960 Mb. The BAC library was pooled and screened by polymerase chain reaction (PCR) amplification using both white clover microsatellites and PCR-based markers derived from *Medicago truncatula*, resulting in an average of 6 hits per marker; this supports the estimated 3-fold genome coverage in this allotetraploid species. PCR-based screening of 766 clones with a multiplex set of chloroplast primers showed that only 0.5% of BAC clones contained chloroplast-derived inserts. The library was further evaluated by sequencing both ends of 724 of the clover BACs. These were analysed with respect to their sequence content and their homology to the contents of a range of plant gene, expressed sequence tag, and repeat element databases. Forty-three microsatellites were discovered in the BAC-end sequences (BESs) and investigated as potential genetic markers in white clover. The BESs were also compared with the partially sequenced genome of the model legume *M. truncatula* with the specific intention of identifying putative comparative-tile BACs, which represent potential regions of microsynteny between the 2 species; 14 such BACs were discovered. The results suggest that a large-scale BAC-end sequencing strategy has the potential to anchor a significant proportion of the genome of white clover onto the gene-space sequence of *M. truncatula*.

Key words: white clover, *Trifolium repens*, *Medicago truncatula*, legumes, comparative genomics, microsynteny, BAC library, BAC-end sequencing, microsatellite.

Résumé : Le trèfle blanc (*Trifolium repens* L.) est une légumineuse fourragère largement cultivée en combinaison avec des graminées en raison de sa capacité à fixer l'azote. Les auteurs ont produit une banque de clones dans des chromosomes bactériens artificiels (BAC) à partir d'une lignée de sélectionneur du trèfle blanc. La banque compte 37 248 clones dont la taille moyenne est d'environ 85 kb, ce qui correspond à environ trois génomes du trèfle blanc sur la base d'un génome estimé à 960 Mb. La banque de BAC a été groupée en pools et criblée par amplification PCR à l'aide de microsatellites du trèfle blanc et de marqueurs PCR provenant du *Medicago truncatula*. En moyenne, six clones ont été obtenus pour chaque marqueur, ce qui est conforme à une couverture équivalent à trois génomes chez cette espèce allotétraploïde. Le criblage par PCR de 766 clones à l'aide d'un multiplexe d'amorces chloroplastiques a révélé que seulement 0,5 % des clones contenaient des inserts dérivés de l'ADN chloroplastique. La banque a été caractérisée plus en détail en séquençant les 2 extrémités de 724 clones BAC. Ces séquences ont été analysées pour ce qui est de leur composition et de leur homologie avec les séquences déposées dans une banque de données comprenant divers gènes, EST et séquences répétées de plantes. Quarante-trois microsatellites ont été décelés parmi les extrémités séquencées (BES) et examinés en tant que marqueurs génétiques potentiels chez le trèfle blanc. Les BES ont également été comparés avec le génome partiellement

Received 26 September 2006. Accepted 10 December 2006. Published on the NRC Research Press Web site at genome.nrc.ca on 31 May 2007.

Corresponding Editor: G.J. Scoles.

M. Febrer. Teagasc, Crops Research Centre, Oak Park, Carlow, Ireland; Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion SY23 3DA, UK.
F. Cheung and C.D. Town. The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA.
S.B. Cannon. USDA-ARS and Department of Agronomy, Iowa State University, Ames, IA 50011, USA.
N.D. Young. Department of Plant Pathology, 495 Borlaug Hall, University of Minnesota, St. Paul, MN 55108, USA.
M.T. Abberton. Legume Breeding and Genetics Team, Institute of Grassland and Environmental Research, Plas Gogerddan, Aberystywth, Ceredigion SY23 3EB, UK.
G. Jenkins. Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion SY23 3DA, UK.
D. Milbourne.[1] Teagasc, Crops Research Centre, Oak Park, Carlow, Ireland.

[1]Corresponding author (e-mail: dan.milbourne@teagasc.ie).

 doi:10.1139/G07-013

séquencé de la légumineuse modèle *Medicago truncatula* dans le but spécifique d'identifier de potentiels BAC représentant des régions de microsynténie entre les deux espèces et 14 de ces BAC ont été identifiés. Les résultats suggèrent qu'une approche de séquençage à grande échelle des extrémités de BAC rendrait possible l'ancrage d'une portion significative du génome du trèfle blanc à l'espace génique du *M. truncatula*.

Mots-clés : trèfle blanc, *Trifolium repens*, *Medicago truncatula*, légumineuses, génomique comparée, microsynténie, banque de BAC, séquençage des extrémités de BAC, microsatellite.

[Traduit par la Rédaction]

Introduction

White clover (*Trifolium repens*) is a perennial forage legume typically found in pasture mixtures with grasses. The advantage of white clover in grass-based forage is increased amounts of nitrogen relative to pure-grass forage. High nitrogen content in white clover is the result of nitrogen fixation by nodule-forming symbiotic bacteria in the plant roots.

Over the past 20 years, much genome-based research has focused on model organisms to define the genetic architecture underlying key processes in a wide variety of organisms. For plants, *Arabidopsis thaliana* was chosen as the model for the dicotyledons because of its small genome size (125 Mb) (Meinke et al. 1998), and rice (*Oryza sativa*) was the first cereal to be sequenced (Nagamura et al. 2003; Zhao et al. 2004). However, in recent years investigators have sought a legume species that could serve as a functional genomic model for certain developmental systems that cannot be studied in *Arabidopsis* (Cook et al. 1997). Two model legumes, *Lotus japonicus* and *Medicago truncatula*, have been adopted to understand the genetic architecture of legume species and to facilitate isolation and characterization of the genes responsible for legume-specific phenomena, including plant–microbe interactions and symbiotic nitrogen fixation (Nakamura et al. 2002).

An important aspect of transferring information from model to agronomically important species is knowledge of the extent of synteny between the two. Within the legumes, the extent of synteny has already been established to varying degrees between *M. truncatula*, *M. sativa* (alfalfa), *L. japonicus* (Choi et al. 2004*a*), *Glycine max* (Yan et al. 2004), and *Pisum sativa* (pea). The extent of similarity between clover and these species remains largely unknown, although preliminary data available at the time of writing suggest that it is quite high. For example, George et al. (2006) discovered a number of microsatellite-containing expressed sequence tags (ESTs) from white clover that had significant matches with the current *M. truncatula* and *L. japonicus* genome sequences, allowing the development of a comparative map that suggests high levels of macrosynteny between clover and *M. truncatula*. In another study, Febrer et al. (unpublished results) tested a set of 100 *M. truncatula* PCR-based markers (originally developed by Choi et al. 2004*b*) in white clover and found a high level of cross-amplification (73%), thereby confirming the sequence-based similarity between the 2 species.

White clover is a member of the tribe Trifolieae, which contains, among others, the closely related genera *Trifolium*, *Medicago*, and *Melilotus*. Although white clover is a self-incompatible allotetraploid and *M. truncatula* is a self-fertile diploid, both are thought to have the same basic number of chromosomes (x = 8), and it is possible that gene order and organisation is largely conserved between the 2 species. Thus, the genome of *M. truncatula*, which is currently being sequenced (Young et al. 2005), may be useful as a tool to identify and isolate orthologous genes in the genome of white clover.

The aim of this work was to construct a bacterial artificial chromosome (BAC) library of white clover for use as a resource to assess the levels of fine-scale conservation of gene order (microsynteny) between this species, *M. truncatula*, and other model genomes. In addition to the construction and basic characterization of the library, we sequenced both ends of over 700 BACs randomly chosen from the library. We present an analysis of these sequences with specific emphasis on the potential of the library for our proposed strategy to "tile" the genome of white clover onto the genome sequence of *M. truncatula*. Our results suggest that, despite evidence for considerable rearrangements between the genomes of white clover and *M. truncatula*, large-scale BAC-end sequencing of the former has the potential to allow the anchoring of a significant portion of the genome of white clover onto that of the latter, significantly improving the potential for gene discovery in white clover.

Materials and methods

BAC library construction

Plant material and isolation of high molecular weight DNA

The white clover mapping parent R3R4, grown in the Institute of Grassland and Environmental Research, was used as the source of high molecular weight (HMW) DNA, derived from isolated nuclei. Cuttings were transferred into 10 cm pots of sterile soil and grown in the glasshouse until 20–30 cm tall. For each round of nuclear isolation, 2–3 selected plants (not more than 4 weeks old) were kept in the dark for at least 48 h to decrease the amount of starch present. Nuclei were isolated from 10 g of leaf material using an adaptation of the method described by Liu and Whittier (1994). Briefly, the leaves were chopped up, swirled in diethyl ether, washed 3 times in sterile water, and added to an extraction buffer containing 100 mL Honda buffer (3.4% Ficoll PM400, 6.8% dextran T40, 34 mmol/L Tris–HCl, pH 8.0, 7 mmol/L $MgCl_2$, 628 mmol/L sucrose, 0.69 mmol/L spermidine, 2.7 mmol/L spermine, and 3.41% *v*/*v* Triton X-100). The mixture was blended by 2 × 5 *s* pulses. The resulting suspension was gravity-filtered through 4–8 layers of Miracloth into a sterile container. The filtrate was centrifuged at 1800*g* at 4 °C for 15 min. The pellet was resuspended in Yamaha buffer (Honda buffer without dex-

tran T40 and Ficoll PM400). If the supernatant was slightly green, the pellet was resuspended and the previous step repeated. The cleaned pellet was then resuspended in a final volume of 400 μL Yamaha buffer without the Triton X-100 and equilibrated at 50 °C for 10 min, after which 500 μL of 1.2% low melting point agarose (also at 50 °C) was added. The mixture was transferred into 100 μL plug moulds and placed on ice. To lyse the embedded nuclei, plugs were incubated at 50 °C with gentle agitation in 50 mL of lysis buffer (0.5 mol/L EDTA, pH 8, 1% *N*-laurylsarcosine, and 0.1 mg/mL proteinase K) for 48 h, with several buffer changes. Finally, plugs were washed twice at 50 °C and once at room temperature in 50 mL of 0.5 mol/L EDTA, pH 8, before being stored in the same buffer at 4 °C.

Partial digestion and size fractionation of HMW DNA

HMW DNA embedded in agarose plugs was incubated for an hour in 50 mL of 1× TE (10 mmol/L Tris-HCl, 1 mmol/L EDTA, pH 8.0) and 0.1 mmol/L phenylmethylsulphonyl fluoride and was washed 4 times for 30 min in 1× TE. This step was repeated twice. Four 100 μL plugs were chopped to a fine granular suspension and placed in a 1.5 mL tube with a TE/Triton X-100 solution (1× TE, 0.1% Triton X-100). The agarose was resuspended by vortexing for 5 s and was then subjected to a 5 s pulse on a benchtop microcentrifuge at full speed. The liquid phase was removed and the resulting granular slurry transferred to six 1.5 mL tubes, in 50 μL aliquots. To each tube was added 15 μL of a buffer containing 7 μL of the appropriate 10× restriction enzyme buffer, 0.7 μL BSA (10 mg/mL), and 7.3 μL of 40 mmol/L spermidine, after which the tubes were equilibrated for 30 min on ice. For partial digestion, 5 μL of diluted *Hin*dIII enzyme was added to each tube to obtain concentrations of 4, 2, 1, 0.5, 0.25, and 0 units of *Hin*dIII per tube. After the addition of enzyme, the mixture was equilibrated on ice for 30 min and then incubated at 37 °C for 30 min, and the reaction was stopped by the addition of 7 μL of 0.5 mol/L EDTA followed by 30 min on ice. The partial digests were loaded on a 1% agarose/1× TBE gel and separated by pulsed-field gel electrophoresis (PFGE) using a CHEF gel apparatus at 14 °C for 18 h (20 s initial switch time, 20 s final switch time, 6 volts/cm, 120° angle). After electrophoresis, the optimal enzyme concentration giving the brightest smear of digested DNA was used for a mass digestion of 4 plugs, carried out as previously described. The gel slice containing DNA ranging in size from 100 kb to 300 kb was excised and submitted to a second size selection to eliminate any small trapped DNA fragments. To compress the DNA into a tighter band on the gel, the gel slice was embedded in the second size selection gel in the reverse orientation. HMW DNA from this gel was excised and electroeluted using Quick-Pik® electroelution capsules (Stratagene). Each 50 μL of eluted DNA containing approximately 100 ng of white clover DNA was used directly for ligation.

Library construction

For ligation, approximately 100 ng of size-selected white clover HMW DNA and 25 ng of the vector pIndigoBAC-5 (*Hin*dIII Cloning-ready, Epicentre) were mixed and incubated with 6 units of T4 ligase (Promega Corporation) at 16 °C overnight. The enzyme was then denatured at 65 °C for 30 min. Ligations were aliquoted into 5 μL, flash frozen in liquid nitrogen, and kept at –80 °C for long-term storage.

Each 5 μL of ligation roduct was transformed into 20 μL of electrocompetent cells (ElectroMax™ DH10B™, Invitrogen) using the Electroporator 2510 (Eppendorf). The transformation mixture was transferred to 1 mL of SOC (20 g/L peptone, 5 g/L yeast extract, 0.5 g/L NaCl, 2.5 mmol/L KCl; 10 mmol/L $MgCl_2$ and 20 mmol/L glucose added prior to use) and incubated at 37 °C for 1 h. Transformed cells were plated onto X/I/C plates (20 g/L Luria Bertani (LB), 1.2% *w/v* technical agar, 90 mg/mL 5-bromo-4-chloro-3-indol-beta-D-galactosidase (X-Gal), 90 mg/mL isopropylthiogalactosidase, 12.5 μg/mL chloramphenicol (Cm)) and incubated at 37 °C overnight. White recombinant colonies were picked directly to 384-well plates (Genetix) containing 60 μL freezing media (2.5% *w/v* granulated LB agar (EM Science), 13 mmol/L KH_2PO_4, 36 mmol/L K_2HPO_4, 1.7 mmol/L sodium citrate, 6.8 mmol/L $(NH_4)_2SO_4$, 4.4% *v/v* glycerol, 12.5 μg/mL Cm). Following overnight incubation at 37 °C, the plates were replicated twice and stored at –80 °C.

Characterization of the BAC library

For insert isolation, BAC clones were inoculated in 3 mL of LB medium with 12.5 μg/mL Cm and grown at 37 °C for 16–18 h. BAC DNA was prepared by an alkaline lysis method (Birnboim and Doly 1979). The final pellet was resuspended in 35 μL of 1× TE, and 5 μL was digested with *Not*I (Sigma) at 37 °C for 1 h. The digested DNA was loaded on a 1% agarose gel and separated by PFGE in 0.5× TBE using a CHEF gel apparatus at 14 °C for 16 h (20 s initial and final switch time, 6 volts/cm, 120° angle). The insert size of BAC clones was determined by comparison with a PFGE lambda ladder (New England BioLabs).

Screening for chloroplast contamination was performed by PCR using consensus chloroplast primer pairs (Chung and Staub 2003). Three primer pairs, ccSSR6, ccSSR15, and ccSSR22, were chosen to screen a set of randomly selected BAC clones. These consensus chloroplast simple sequence repeats are contained in genes based in the tobacco chloroplast genome (*RpoB*, *Rpl20-ClpP*, *TrnL-16SrRNA*, respectively) and are evenly distributed across it (position according to the tobacco chloroplast genome, accession number CHNTXX). The BAC clones were incubated in a 384-well plate containing 10 μL LB with 12.5 μg/mL Cm and grown at 37 °C overnight, and 1 μL of each culture was added to a 10 μL multiplex PCR reaction (0.3 mmol/L deoxynucleoside triphosphates (dNTPs), 2 mmol/L $MgCl_2$, 1× buffer, 30 ng of each primer, 0.2 units *Taq* polymerase). The amplification conditions were 94 °C, 5 min; 35 cycles of 94 °C for 1 min, 50 °C for 1 min, 72 °C for 1 min; and finally 72 °C, 6 min. After PCR the products were analysed in a 2% agarose gel and the chloroplast contamination was determined by the frequency of occurrence of the expected amplicons for the primer sets.

To prepare pools of BAC DNA for PCR amplification screening, each library plate was replicated onto a Petri dish with LB agar containing 12.5 μg/mL Cm, using 384-pin manual plastic replicators (Genetix). After overnight incubation at 37 °C, arrayed colonies from each Petri dish were pooled by the addition of 5 mL LB medium, into which the

colonies were scraped from the surface of the LB agar. For each plate pool, 2 mL of colony suspension was subjected to alkaline lysis to recover BAC DNA. The plate pools were then screened by PCR amplification with white clover microsatellites (SSRs) (Barrett et al. 2004) and PCR-based markers from *M. truncatula* that had previously been shown to produce single-copy amplicons in clover (data not shown). Four SSRs (ATS123, PRS 256, ATS176, ATS055) were tested, and PCR was set up using 2.5 ng of plate-pooled DNA transferred into 18 μL of reaction mix (1× buffer, 0.2 mmol/L dNTPs, 0.2 μmol/L primers, 0.06 units *Taq* polymerase) under amplification conditions of 94 °C, 4 min; 30 cycles of 94 °C for 30 s, 55 °C for 30 s, and 72 °C for 1 min; then 72 °C, 7 min. A slightly altered method was used for the PCR-based markers (DK501R, DNABP, TR59, TR66, TR68), whereby 2.5 ng of plate-pooled DNA was transferred into 18 μL of reaction mix (1× buffer, 0.2 mmol/L dNTPs, 1 mmol/L $MgCl_2$, 0.1 μmol/L primers, 0.05 units *Taq* polymerase) and amplification conditions were 94 °C, 3 min; 35 cycles of 94 °C for 45 s, 55 °C for 45 s, 72 °C for 1.30 min; then 72 °C, 7 min.

BAC-end sequencing

DNA template was prepared in 384-well format by a standard alkaline lysis method. End sequencing was performed using Applied Biosystems (ABI) Big Dye Terminator chemistry and analysed on ABI 3730x1 machines. Base calling was performed using TraceTuner, and sequences were trimmed for vector and low-quality sequences using LUCY (Chou and Holmes 2001). Sequences were compared with all entries in The Institute of Genomic Research (TIGR) Plant Gene Indices (http://www.tigr.org/plantProjects.shtml) using BLASTn with a cut-off value of 1×10^{-20} (Quackenbush et al. 2000); a BLASTx (cut-off value 1×10^{-10}) was used in the TIGR nonidentical amino acids database, which contains nonidentical protein data from GenBank, RefSeq, Uniprot, Comprehensive Microbial Resource, Protein Data Bank, and Protein Research Foundation. The BAC-end sequences were also compared with repetitive DNA in the TIGR transposon database using BLASTx with a cut-off value of 1×10^{-10}. The BAC-end sequences were compared with the 167 690 648 bp of *Medicago* genome sequence contig data available on 12 September 2005 (http://www.tigr.org/tdb/e2k1/mta1/) using BLASTn with a cut-off value of 1×10^{-10}. To identify comparative-tile BACs from the clover library that were likely collinear (i.e., showed microsynteny) with the *M. truncatula* genome, the searches against the *Medicago* genomic sequence were parsed first to remove transposon matches and then to identify BACs for which both ends had a significant match to a stretch of *Medicago* sequence and for which the 2 regions on the *Medicago* genome were between 20 kb and 200 kb apart.

Development and mapping of microsatellites from the BAC-end sequencing

The BAC-end sequences (BESs) containing SSRs were identified using the MISA (microsatellite identification tool, Thiel et al. 2003). Primers for BESs containing SSRs were designed using Primer 3 software. Each primer pair was tested for amplification by PCR on the genotype used for the BAC library construction (R3R4) and another genotype (S1S4), the parents of a reference mapping family with 94 progeny. The primer pairs were then tested for polymorphism using radioactive labelling in the 2 mapping parents and 4 F_1 progeny. The PCR was set up in a final volume of 10 μL as follows: 1.25 ng genomic DNA, 0.025 units *Taq* polymerase (New England Biolabs), 1× PCR buffer (New England Biolabs), 0.2 mmol/L dNTPS, 0.25 μmol/L radioactively labelled forward primer (using [^{33}P]-ATP from GE Healthcare), and 0.1 μmol/L of the reverse primer. PCR conditions consisted of 3 min at 94 °C followed by 34 cycles of 45 s at 94 °C, 45 s at 50 °C, and 45 s at 72 °C, and a final extension step of 2 min at 72 °C. The PCR products were analysed on a 5% polyacrylamide gel, and the gels were dried on Whatmann 3MM paper and exposed to storage phosphor screens for 1 to 3 d at room temperature.

All polymorphic SSRs were subsequently analysed in the entire population using the ABI 3100 sequencer (Applied Biosystems). Each marker was labelled with a different fluorophor (FAM = blue, NED = yellow, VIC = green, PET = red), allowing the PCR products to be pooled. The reaction for each primer was set as follows: 2.5 ng/μL DNA, 625 μmol/L dNTPs, 1 unit *Taq* polymerase, and 0.25 μmol/L of forward and reverse primers (NED, PET, and VIC, Applied Biosystems) or 0.5 μmol/L of forward and reverse primers (FAM, MWG Biotech, Germany). The PCR program was as follows: denaturation at 95 °C for 5 min; amplification for 36 cycles of 95 °C 1 min, 55 °C 1 min, and 72 °C 1 min; and extension at 72 °C for 10 min. The plates were then incubated at 60 °C for 30 min to avoid the formation of polyadenosine peaks. The PCR products were then pooled as follows: 1 μL FAM, 1 μL VIC, 2 μL NED, and 2 μL PET. On an ABI 96-well plate, 0.5 μL of the pooled PCR products was added to 9.5 μL of formamide/sizer (25 μL size standard (GeneScan™, 500 LIZ® Size Standard) + 950 μL formamide (Hi-Di™ Formamide, ABI)). The plate was incubated at 95 °C for 5 min and placed immediately on ice. It was then run on the ABI 3100 Genetic Analyser and the data were analysed using ABI PRISM® GeneMapper™ software Version 3.0.

SSRs analysed as described earlier were scored and subsequently incorporated into the reference genetic linkage maps of the genotypes S1S4 and R3R4 using JoinMap® 3.0 (Van Ooijen and Voorrips 2001).

Results

Construction of the BAC library

A BAC library of mapping parent R3R4 of white clover was constructed from HMW DNA isolated from nuclei embedded in agarose plugs using an adapted method from Liu and Whittier (1994). HMW DNA was released from agarose by electroelution prior to partial digestion with *Hin*dIII. According to PFGE analysis, 0.5 units of *Hin*dIII per 50 μL of plug solution produced the largest amount of DNA in the 50–200 kb size range (data not shown). The library was constructed with 2 size selections of partially digested DNA to promote the release of any small, trapped DNA fragments. Five separate ligations gave rise to the library consisting of 37 248 clones arrayed in 97 × 384-well plates.

Characterization of the library

To determine the average insert size of the library, DNA from 86 randomly selected BAC clones was isolated and digested with *Not*I enzyme to release the DNA insert from the cloning vector. Out of the 86 BAC clones tested, all contained white clover insert DNA. The insert sizes ranged from 45 to 146 kb with an average of 85 kb (Fig. 1). The size distribution of these clones (Fig. 2) approximates a normal distribution, with the majority of clones in the 80 to 90 kb size range. On the basis of an estimated genome size of 960 Mb for white clover, the library provides coverage of approximately 3 genome equivalents.

To obtain an estimate of the representation of chloroplast DNA in the library, a set of randomly selected clones were screened against a set of consensus chloroplast primers (Chung and Staub 2003) evenly distributed at intervals of approximately 50 kb across the chloroplast genome, which is below the average insert size of the library. A multiplex PCR amplification performed on 766 BAC clones showed that 0.5% of the library contains chloroplast DNA.

Preparation of BAC DNA pools and PCR amplification screening of the library

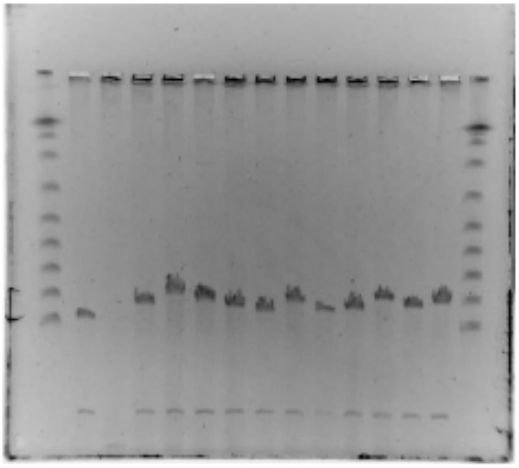
To confirm the estimated genome coverage of the BAC library, PCR-based screening of plate pools was performed using white clover SSR and PCR-based markers from *M. truncatula* chromosome 3. BAC DNA from each plate of the library was pooled, resulting in 97 plate pools. The BAC DNA plate pools were screened by PCR using 4 white clover SSR primer pairs (ATS123, PRS256, ATS176, ATS055) previously mapped in a white clover mapping population of which R3R4 was one of the parents, and 5 PCR-based markers from *M. truncatula* (DK501R, DNABP, TR59, TR66, TR68) that have previously been shown to produce single-copy amplicons in white clover (data not shown). Screening with these markers resulted in a range of 1 to 9 hits, with an average of 5.7 hits per marker (SD 2.65). Given the allotetraploid nature of white clover, each single-copy marker having 2 potential homoeologues in the genome and assuming that each hit in a BAC pool represents only 1 positive BAC in that pool, an average of approximately 6 hits per marker is completely consistent with the estimated 3-fold genome coverage of the library.

BAC-end sequencing

When an extensively sequenced model species exists that is closely related to a relatively uncharacterized species, high-throughput BAC-end sequencing offers the potential to "tile" the genome of the uncharacterized species onto that of the sequenced species. To further characterize our BAC library and to test whether a BAC-end sequencing approach might be effective for white clover in the manner already described (using *M. truncatula* as the reference species), the contents of 2 × 384-well plates were end-sequenced, resulting in 1474 high-quality BESs and a total of 725 BES pairs. The average vector-trimmed read length of the BESs was 800 bp with a total read length of 1.61 Mb. All sequences have been submitted to GenBank, with the accession numbers ED549329–ED550789.

Comparison of the 1474 BESs with the TIGR Plant Gene Indices and the TIGR nonidentical amino acids database revealed that 368 of the sequences (24.9%) could be identified as "genic" in nature by virtue of good matches to either ESTs (BLASTn, cut-off 1×10^{-20}) or representative protein-coding sequences from the nonredundant amino acid database (BLASTx, cut-off 1×10^{-10}) (Table 1). Of the 368 genic sequences, the top BLAST match in 258 cases was to a legume EST. Of these 258 legume matches, 69.4% (179) were to *Medicago* ESTs, 22.1% (57) to soybean ESTs, and 8.5% (22) to *Lotus japonicus* ESTs (Table 2). In total, 251 of the 368 genic sequences (68.2%) had a good BLAST match (above the cut-off of 1×10^{-10}) to a *Medicago* EST, reflecting the high level of relatedness between the 2 species.

Fig. 1. An analysis of white clover BAC clones by PFGE. Lanes 1 and 15, lambda ladder (bands increasing in 50 kb increments, starting at 50 kb); Lanes 2–14, fragments of *Not*I-digested DNA isolated from randomly selected BAC clones.



Of the 1474 BAC-end sequences analysed, 126 (8.5%) were found to contain sequences homologous to transposable elements (cut-off 1×10^{-10}) (Table 1). The majority of transposable elements belonged to the Ty3_copia type (68.3%) followed by the Ty1_gypsy (23.8%) and LINE (7.9%) types of retrotransposons. In comparison, 16.6% of 165 643 *Medicago* BAC ends were homologous to transposable elements (cut-off 1×10^{-10}). We do not yet know, however, to what extent the higher proportion of transposable elements identified in *Medicago* BAC ends (16.6% vs. 8.5% in clover) was affected by the presence of large numbers of *Medicago* transposable elements in the target database.

When the 1474 white clover BESs were compared with the *M. truncatula* genome sequence, 57% (844) of them had a significant hit (1×10^{-10}) to *M. truncatula* BACs, with percentage identities ranging from 78% to 100% for top matches. Comparisons were made with all *M. truncatula* (Phase I, II, and III) BAC sequences available at the time of writing. Of the 1474 BAC-end sequences, 1450 were paired sequences in which the sequence at the other end of the BAC was also known (giving a total of 725 BES pairs). Of the 725 BES pairs, 204 had a significant BLAST match (both members of the pair) to *M. truncatula* genome se-

Fig. 2. Distribution of insert sizes of randomly selected BAC clones.

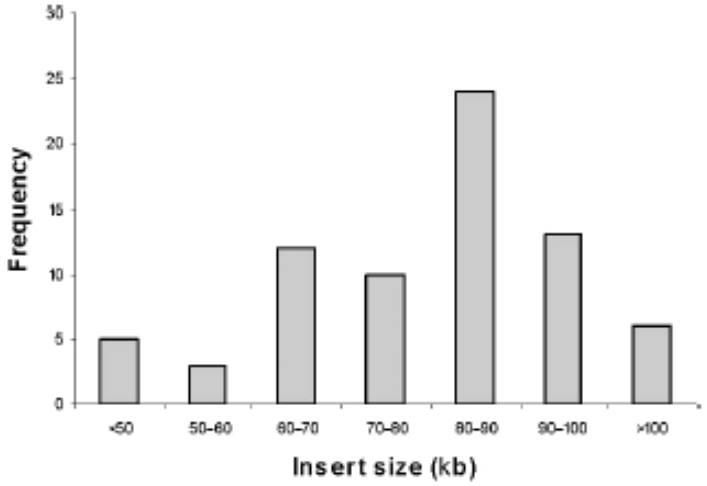


Table 1. Summary of the content and composition of the BAC-end sequences.

Number of BAC clones sequenced	768
Number of BAC-end sequences	1474
Number of paired BAC ends	725
Average read length (bases)	800
Total sequence length (Mb)	1.16
Sequence composition	
Genic/protein coding (%)	24.9
Transposable elements (%)	8.5
Microsatellites (%)	2.9

Note: BAC, bacterial artificial chromosome.

Table 2. EST hit summary.

Species	No. of hits
Legumes	
Medicago	179
Soybean	57
Lotus japonicus	22
Other	
Arabidopsis	26
Rice	25
Grape	10
Cotton	6
Chlamydomonas reinhardtii	6
Saccharum officinarum	6
Tomato	5
Potato	4
Tobacco	4
Aquilegia	4
Maize	3
Barley	3
Sorghum	2
Poplar	2
Lettuce	2
Beet	1
Pepper	1
Total BESs	1474
No. of EST hits	368

Note: BES, bacterial artificial chromosome end sequence; EST, expressed sequence tag.

quence. Of these, 14 were shown to have the equivalent pairs of *M. truncatula* sequence on the same *M. truncatula* BAC clone or contig within a span of 20 to 200 kb. Clover BACs that fulfil these criteria are putative comparative-tile BACs and potentially represent regions of highly conserved gene content and organisation between clover and *M. truncatula*. The *Medicago* matches to the 14 paired clover BAC ends ranged in distance from 23.1 kb to 121.7 kb (Table 3). The size of the clover BACs (and thus the distance between the end sequences) was compared with the span by which the paired matches were separated in the *Medicago* genome (Table 3). In 11 of the 14 cases, the separation span in clover exceeded that in *Medicago* (with differences ranging from 18.8 kb to 70.9 kb), and in 3 cases the span in *Medicago* exceeded that observed in clover (with differences ranging from 27.9 kb to 64.6 kb).

Among the BAC-end sequences, 43 novel microsatellites of a length sufficient for potential marker development were identified. The predominant motif was the dinucleotide repeat (58.12%), followed by tri (20.9%), tetra (11.6%), and mononucleotides (2.3%) with only 3 (7%) compound microsatellites. The number of repeats in the microsatellite motifs ranged from 5 to 34. It was possible to design primer sets to 34 of these SSRs (Table 4). All of the primers were tested in a small set of clover germplasm comprising the R3R4 genotype from which the library was constructed, another genotype referred to as S1S4, and a number of F_1 progeny from a mapping population of which these 2 genotypes are the parents (Fig. 3). This population has been used at Oak Park to construct a map, which comprises 474 AFLP (amplified fragment

Table 3. The 14 paired BAC ends that had paired BLAST matches in *M. truncatula* genome sequence at a distance of between 25 kb and 200 kb.

BES	Accession No.	Size in *Medicago* (kb)	Size in white clover (kb)
WCBE306TF	ED549888	51.8	72
WCBE306TR	ED549889		
WCBE216TF	ED549723	40	86
WCBE216TR	ED549724		
WCBE144TF	ED549591	67	104
WCBE144TR	ED549592		
WCBE053TF	ED549426	49.6	93
WCBE053TR	ED549427		
WCBE349TF	ED549969	115.4	82
WCBE349TR	ED549970		
WCBE166TF	ED549632	51.6	114
WCBE166TR	ED549633		
WCBE368TF	ED550005	81.2	100
WCBE368TR	ED550006		
WCBE181TF	ED549662	121.6	57
WCBE181TR	ED549663		
WCBE088TF	ED549496	46.3	89
WCBE088TR	ED549497		
WCBE614TF	ED550451	40.4	95
WCBE614TRB	ED550452		
WCBE732TF	ED550665	23.1	94
WCBE732TRB	ED550666		
WCBE735TF	ED550671	92.8	114
WCBE735TRB	ED550672		
WCBE546TF	ED550325	102.9	75
WCBE546TRB	ED550326		
WCBE748TF	ED550697	55.6	107
WCBE748TRB	ED550698		

length polymorphism) markers and 46 SSR markers, and covers all 16 expected linkage groups of white clover, with a number of short, unassigned linkage groups (map not shown). All of the primers that were tested successfully amplified products of the expected size. Of the 34 primer pairs tested, 21 (61%) revealed a polymorphic pattern that was reliably scoreable over the entire mapping population, which is in keeping with the levels of SSR polymorphism previously observed in this population (data not shown). On analysis in JoinMap® 3.0, 18 of the SSRs were successfully incorporated into the genetic map. The genetic map is not shown, but the clover linkage groups that the SSRs mapped to are listed in Table 4 (using the linkage group nomenclature of Barrett et al. 2004), along with the accession number of the BAC from which they were derived and the primer sequences.

Discussion

We have constructed a 3-genome equivalent BAC library of white clover mapping parent R3R4 using partial *Hin*dIII digests. The library consists of 37 248 clones with an average insert size of approximately 85 kb. Taking into account that 0.5% of the library clones contain chloroplast DNA, the library comprises a total of approximately 3000 Mb of nuclear DNA. This equates to a 95.8% probability of containing any sequence in the white clover genome (Clarke and Carbon 1976). The method used for nuclei extraction in this study, while similar in principle to methods used previously to construct BAC libraries, was adapted from a technique used to extract crude nuclear proteins in gel mobility shift assays, and to our knowledge this is its first reported use for BAC library construction. It was found to be very successful in white clover in terms of the small amount (10 g) of leaf material needed to produce HMW DNA of sufficiently high quality for the construction of a 3× library, with reasonably low levels (<1%) of chloroplast contamination. Screening of the BAC library with both white clover microsatellites and PCR-based markers from *M. truncatula* resulted in the detection of an average of ≈6 clones per marker, consistent with the calculated genome coverage of the library representing 3 haploid-genome equivalents.

The analysis of BAC clones by *Not*I digestion followed by PFGE showed that the majority of white clover DNA inserts were present as single *Not*I fragment inserts, implying that the white clover genome apparently contains few *Not*I sites. This reflects the low level of GC richness, which is a typical feature of dicot genomes in comparison to monocots (Choi et al. 1995; Danesh et al. 1998; Tomkins et al. 1999; Meksem et al. 2000). All the 86 randomly chosen clones tested contained an insert, implying a low percentage of empty clones in the library.

Comparison of the clover BESs with both EST database entries and the available genome sequence data for *M. truncatula* lends further support to the proposed high level of sequence-based similarity between these 2 legume species. Out of the 368 clover BESs that were classed as genic in nature, ≈70% had a significant BLASTn match to *Medicago* ESTs, a figure that is in keeping with the rate of cross-species amplification of *M. truncatula*-derived PCR-based markers in clover mentioned earlier. When compared with the *M. truncatula* genome sequence, 57% (844) of the 1474 clover BESs had a good BLAST match. Taking into account potentially spurious matches involving repeat elements, ≈50% of the clover end sequences had a good BLAST match in the *Medicago* genome sequence. Given that the comparison is based on approximately 168 Mb of *Medicago* sequence contig data, which corresponds to ≈34% of the entire genome of *M. truncatula*, the proportion of clover BACs with good *Medicago* matches is actually relatively high. In addition, the high levels of similarity extend beyond those clover BESs that we defined as genic by virtue of either a protein or an EST match, as 649 of the clover BES that had a good match in *M. truncatula* were neither genic (as defined earlier) nor repeat-element-based.

A major goal in comparing our clover BES data to the *Medicago* genome sequence data was to gain a preliminary insight into the extent of microsynteny between the 2 species and to investigate the utility of a large-scale BES strategy to "tile" a significant proportion of the genome of white clover onto that of *Medicago*. The latter would constitute a useful translational genomics platform for gene isolation in white clover in the absence of significant amounts of publicly available sequence information in this species. To that end, we identified all clover BACs that had a significant BLASTn hit to *Medicago* genome sequence at both ends and the subset of these for which both BAC ends had good

Table 4. SSRs from the BAC-end sequence.

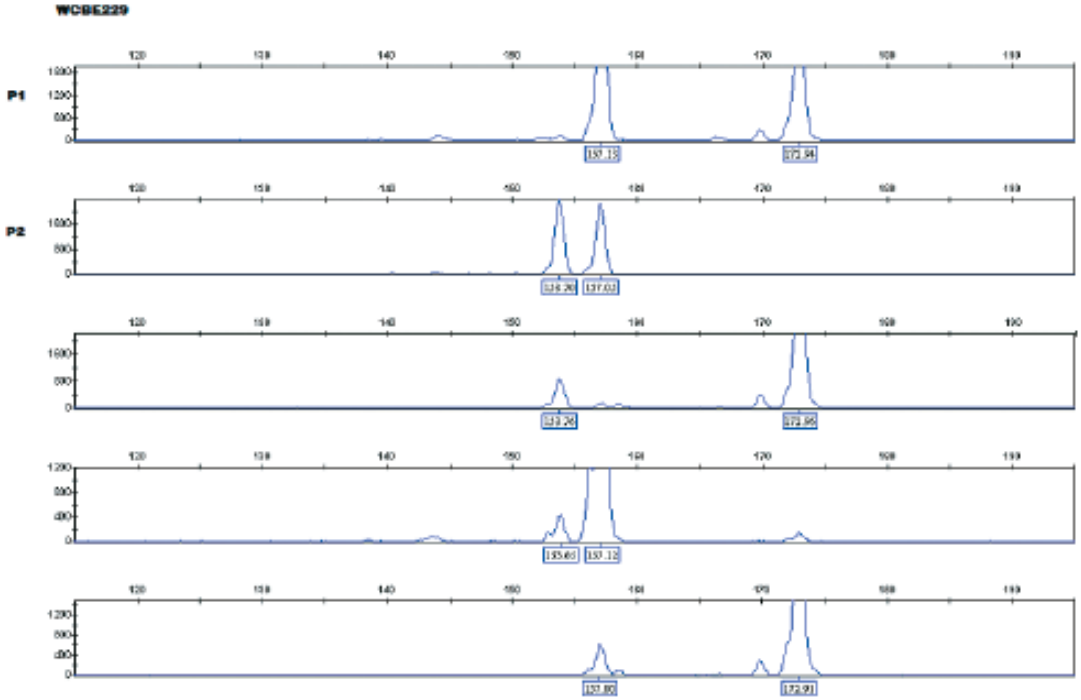
SSR ID	Forward primer	Reverse primer	White clover chromosome location[a]
WCBE005TF	TGCATCTTATCTACGCACTTTTG	CAAAATTAAAGAATCAAACCATGA	D
WCBE010TF	CCGGTGTAATTTGATCTCTCA	AAAAATCAAAGGTCATGATCTAAATG	Unidentified[b]
WCBE028TR	TGAATTGTGAATGAATGAAAA	TCATTCTCTTTCAAAATTTCTCCA	Not polymorphic
WCBE040TF	CGCGAGAATCAACAGAAACC	CACTTGGTGAAGTGGTCCAT	Unidentified
WCBE073TF	TTTTGTGGCGGTTTCTATTG	TTGCCAATTGCCATCATGTA	A
WCBE086TF	TCTGCATTAGGCTTGGTGTG	AAGAAAGGGTCACCACAAAA	Not polymorphic
WCBE113TF	AGTGTTTAATAATTGGTTAAATTTTGG	TTGCATTTTCTAATCTACACAATTTCT	A
WCBE158TR	TTGCGCTTTACCGTTATTCC	CGATGACGACGATGGTTACA	G
WCBE160TF	GGAGTATTTCCCACCCCACT	TTCTACACCCGCGAGAGTTC	F
WCBE209TF	TTGGAGTTATAAATGACAAAATTGC	CTTGCAATTGATCCTACATACG	Not polymorphic
WCBE224TR	TTGCATACTGTGTTTTTCTTTTTGA	TTTTACGCATTGTGCTCGTC	Unidentified
WCBE229TR1	TTCACACCCTATACAATGGAGTAGT	CCCATTTTTCTCAAATCACAAAG	Not polymorphic
WCBE229TR2	TGAACAATAAAGCTTTTGGCTAAG	CAAATGAGTGAATGCCATGA	Unidentified
WCBE307TF	TTTGGATGAATTTATTTTGGGTTT	GCCCCTGGGTAACAAATTCT	G
WCBE325TF	AATCCATTAGCGGACTTCCA	CAGTGTTTAACATCTTGAAACCATTTA	Not polymorphic
WCBE336TF	AAATCCATTTTGCAAGATAACCA	ATCTCAAGCAAAGTAATTTTCAAGG	Not polymorphic
WCBE364TR	GCTCAACCTCCCAAACAACT	CCAACTTTGCAATTAACTCATCA	A, C
WCBE381TR	ACCAAGCAAGAAAGGCAAGA	TGAATAATTTTCTTCATTGTTGTTTAC	Not polymorphic
WCBE387TR	TCACCAAGGGGTAATGGAAC	CAAGTTTGAATTATGGTGGAGTG	Not polymorphic
WCBE468TF	TGATACTGTCGCGATACGTTATTT	GCAAGCATATTGGCTTGTTTG	Not polymorphic
WCBE513TF	TCTGACTCAACATTCTGTTTTCA	AACAAAAACAGGTATCGAACAGC	Unidentified
WCBE517TRB	GCGGCCAATATTGTTTGAAG	GTCGATTTCGAGCAGATTGA	D
WCBE566TRB	GGCGGAGGCTAGGGTAAAATC	AAAGCGCTATTGAGAAGTAT	D
WCBE578TF	GATGGTGGGAACAATTTTGG	ATCCGGAACGATTCGTGTAG	F
WCBE579TF	TCTTTCTTGCATTATGGCTCA	TTTCTGGAGCTCGGTACCAC	Not polymorphic
WCBE582TRB	GCAATTGGATTACCGACGAA	AAAAGCGCGTTCAACAATCT	Not polymorphic
WCBE602TF	ACCACGTGGGATGGAGAGTA	AACTTGCTAATGCACACCTTCA	G
WCBE655TF	CCATTGCGTATATGGAGCTG	TTTTCTTTAATTATTTCTCATGCGTTA	D
WCBE692TF	CACCAGCTGTCATTCCAAAA	GTGTGTGGAAAGGCAATCAA	G
WCBE712TRB	GAACAGGAGTTGGACGCAAT	TTGCCATCGATTTCTGCTAA	B
WCBE747TRB	TTTTGCTGTAAGAACGCATTG	GAGTTCTATATGAATTAATGGGAATCA	Not polymorphic
WCBE773TRB	GGGTTGGATTGGAGAATCTG	GCCATCAAATATCCACACACA	Not polymorphic
WCBE776TRB	CATCTTGGTTTTCCAGAATGG	GCTGGGTTGTTGCCTAACTC	Not polymorphic
WCBE791TRB	GAACTAAATACTGACATGGGCAAT	CAATTGAAAATTTGAAAGGGATCT	E

[a]Chromosomal location according to the naming convention of Barrett et al. (2004). Homoeologous groups (e.g., A1, A2) have not been distinguished.
[b]Maps to a linkage group that does not currently have a chromosomal designation in our map.

BLASTn matches in *Medicago* genome sequence at a distance consistent with the span of a BAC. Of 204 clover BACs with a *Medicago* genome sequence match on both ends (at a threshold of 1×10^{-10}), 14 had matches that were separated by a compatible distance in *Medicago*, and, consistent with other studies of this nature, we propose that these represent regions of conserved microsynteny between white clover and *Medicago*. It is interesting to note that, in actual fact, the majority (190) of the clover BACs with a BLASTn match to *Medicago* on both ends did not conform to our conditions for identity as comparative-tile BACs.

It is not possible to comment extensively on the potential microsyntenic relations between these clover BACs and the *Medicago* genome. Plant genomes are highly repetitive and, in evolutionary terms, subject to segmental duplication events followed by differential gene loss, processes that tend to result in a fragmentary pattern of conserved microsynteny even between relatively closely related plant genomes (Lagercrantz 1998; Bennetzen and Ramakrishna 2002). An additional complication in this case is the fact that white clover is thought to be an allotetraploid descended from a proposed ancestral hybridization between 2 diploid progenitor species. Recent molecular phylogenetic analysis using both chloroplast and nuclear DNA markers supports the possibility that *T. occidentale* and *T. pallescens* were these progenitors (Ellison et al. 2006). While genetic mapping studies support conservation of macrosyntenic relations between the 2 homoeologous genomes of the species, little is really known about the true extent of conserved microsynteny between the genomes. Both of these factors could contribute to the observed effect of a significant number of paired-clover BESs seeming to "map" to completely different parts of the *Medicago* genome. It would be possible to counteract this by applying more stringent parameters (for the identification of putative orthologues between clover BESs and *Medicago* genome sequence). For example, applying a set of filtering criteria to the current data set including a BLASTn threshold of 1×10^{-30}, a minimum percentage identity of 80%, and a total number of *Medicago* matches lower than 30 results in a drop in the initial number of

Fig. 3. ABI chromatogram of one of the BAC-end sequence SSRs (WCBE229). The 2 mapping parents (P1 = R3R4 and P2 = S1S4) are followed by 3 progeny individuals from the population.



match pairs from 204 to 22 (data not shown). Such increased stringency filtering might be quite useful in the context of a high-throughput BES initiative, as described below.

Although the proportion of comparative-tile BACs discovered seems quite low, it does suggest that high throughput BAC-end sequencing of white clover would allow anchoring of a significant portion of the clover genome onto the *M. truncatula* genome sequence. For example, end sequencing 75 000 BACs (equivalent to 6 haploid genome equivalents in a library with our relatively modest insert size) would result in approximately 1400 comparative-tile BACs for white clover, corresponding to 119 Mb of clover genome, not taking into account potential overlapping coverage of the *Medicago* genome by clover comparative-tile BACs. This figure would significantly increase following the sequencing of the remainder of the gene space of *M. truncatula*, of which an estimated 60% is currently sequenced. Thus, it is feasible that a significant fraction of the gene space of white clover could be anchored to the gene space of *M. truncatula* using this approach.

We discovered 43 novel clover SSRs among the BESs and explored their utility as genetic markers in a clover mapping population. In a preliminary study, George et al. (2006) have proposed that each of the 8 homoeologous pairs of clover chromosomes (named A_1, A_2–H_1, H_2) are broadly homologous to 1 of the 8 *Medicago* chromosomes (numbered 1–8). This might lead to the expectation that at least some of the SSR markers derived from clover BESs that have significant BLAST matches to the *Medicago* genome sequence would map to the clover linkage group that is the proposed orthologue of the *Medicago* chromosome on which the top BLAST match for that clover BES was found. Such markers could act to anchor the clover genetic map directly to the *Medicago* genome sequence map. This scenario was not found for any of the 8 mapped clover SSRs in this study, which were derived from BESs with a significant BLAST match in the *Medicago* genome sequence. However, it is worth noting that 6 of the 8 clover BAC-end sequences in question actually had multiple BLAST matches (ranging from 4 to 163 matches at the 1×10^{-10} threshold) to *Medicago* genome sequence, and in 4 of these cases lower significance BLAST matches were located on the *Medicago* chromosome consistent with the proposed orthologous group in clover. These results suggest that while BES-derived microsatellites are a viable source of markers for white clover, it would be problematic to use them in the way suggested earlier to anchor the genome of clover onto that of *M. truncatula*.

In conclusion, we have developed a BAC library of white clover, and preliminary BAC-end sequencing analysis supports a significant retention of synteny between the genomes of white clover and *M. truncatula*, consistent with other findings of extensive synteny across the cool-season legumes (Choi et al. 2004*a*; Choi et al. 2004*b*; Cannon et al. 2006). The BAC library is available as a resource for translational genomics and gene isolation in white clover.

Acknowledgements

Research at Teagasc was funded under the Irish National Development Plan. Research at TIGR and the University of Minnesota was supported by National Science Foundation Award Number DBI-0321460. The Institute of Grassland and Environmental Research is grant aided by the Biotechnology and Biological Sciences Research Council.

References

Barrett, B., Griffiths, A., Schreiber, M., Ellison, N., Mercer, C., Bouton, J., et al. 2004. A microsatellite map of white clover. Theor. Appl. Genet. 109: 596–608. PMID:15103407.

Bennetzen, J.L., and Ramakrishna, W. 2002. Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. Plant Mol. Biol. 48: 821–827. doi:10.1023/A:1014841515249. PMID:11999852.

Birnboim, H.C., and Doly, J. 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. 7: 1513–1523. PMID:388356.

Cannon, S.B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J.P., et al. 2006. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genome. Proc. Natl. Acad. Sci. U.S.A. 103: 14959–14964.

Choi, S.D., Creelman, R.A., Mullet, J.E., and Wing, R.A. 1995. Construction and characterization of a bacterial artificial chromosome library of *Arabidopsis thaliana*. Weeds World, 2: 17–20.

Choi, H.K., Kim, D.J., Uhm, T., Limpens, E., Lim, H., Mun, J.H., et al. 2004*a*. A sequence-based genetic map of *Medicago truncatula* and comparison of marker colinearity with *Medicago sativa*. Genetics, 166: 1463–1502. doi:10.1534/genetics.166.3.1463. PMID:15082563.

Choi, H.K., Mun, J.H., Kim, D.J., Zhu, H., Baek, J.M., Mudge, J., et al. 2004*b*. Estimating genome conservation between crop and model species. Proc. Natl. Acad. Sci. U.S.A. 101: 15289–15294. doi:10.1073/pnas.0402251101. PMID:15489274.

Chou, H.H., and Holmes, M.H. 2001. DNA sequence quality trimming and vector removal. Bioinformatics, 17: 1093–1104. doi:10.1093/bioinformatics/17.12.1093. PMID:11751217.

Chung, S.M., and Staub, J.E. 2003. The development and evaluation of consensus chloroplast primer pairs that possess highly variable sequence regions in a diverse array of plant taxa. Theor. Appl. Genet. 107: 757–767. doi:10.1007/s00122-003-1311-3. PMID:12827249.

Clarke, L., and Carbon, J. 1976. A colony bank containing Col El hybrid plasmids representative of the entire *E.coli* genome. Cell, 9: 91–99. doi:10.1016/0092-8674(76)90055-6. PMID:788919.

Cook, D.R., VandenBosch, K, de Bruijn, F.J., and Huguet, T. 1997. Model legumes get the nod. Plant Cell, 9: 275–281. doi:10.1105/tpc.9.3.275.

Danesh, D., Penuela, S., Mudge, J., Denny, R.L., Nordstrom, H., Martinez, J.P., and Young, N.D. 1998. A bacterial artificial chromosome library for soybean and identification of clones near a major cyst nematode resistance gene. Theor. Appl. Genet. 96: 196–202. doi:10.1007/s001220050727.

Ellison, N.W., Liston, A., Steiner, J.J., Williams, W.M., and Taylor, N.L. 2006. Molecular phylogenetics of the clover genus (*Trifolium*–Leguminosae). Mol. Phylogenet. Evol. 39: 688–705. PMID:16483799.

George, J., Cogan, N.O.I., Smith, K.F., Spangenberg, G.C., and Forster, J.W. 2006. Genetic map integration and comparative genome organization of white clover (*Trifolium repens* L.) with model legume species. Plant and Animal Genomes XIV conference, 14–18 January 2006. Town & Country Convention Center, San Diego, Calif. 452 pp. Available from http://www.intl-pag.org/14/abstracts/PAG14_P452.html

Lagercrantz, U. 1998. Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. Genetics, 150: 1217–1228. PMID:9799273.

Liu, Y.G., and Whittier, R.F. 1994. Rapid preparation of megabase plant DNA from nuclei in agarose plugs and microbeads. Nucleic Acids Res. 22: 2168–2169. doi:10.1093/nar/22.11.2168. PMID:8029028.

Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D., and Koornneef, M. 1998. *Arabidopsis thaliana*: a model plant for genome analysis. Science (Washington, D.C.), 282: 662–682. doi:10.1126/science.282.5389.662. PMID:9784120.

Meksem, K., Zobrist, K., Ruben, E., Hyten, D., Quanzhou, T., Zhang, H.B., and Lightfoot, D.A. 2000. Two large-insert soybean genomic libraries constructed in a binary vector: applications in chromosome walking and genome wide physical mapping. Theor. Appl. Genet. 101: 747–755. doi:10.1007/s001220051540.

Nagamura, Y., Horiuchi, I., Sugioka, K., Watanabe, K., Antonio, B.A., Akimoto, M., et al. 2003. Rice-BLAST: a comprehensive homology search for rice specific sequences. Genome Informatics, 14: 533–534.

Nakamura, Y., Asamizu, E., and Kaneko, T. 2002. A legume *Lotus japonicus* genome annotation. Genome Informatics, 13: 539–540.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. 2000. The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res. 28: 141–145. doi:10.1093/nar/28.1.141. PMID:10592205.

Thiel, T., Michalek, W., Varshney, R.K., and Graner, A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor. Appl. Genet. 106: 411–422. PMID:12589540.

Tomkins, J.P., Yu, Y., Miller-Smith, H., Frisch, D.A., Woo, S.S., and Wing, R.A. 1999. A bacterial artificial chromosome library for sugarcane. Theor. Appl. Genet. 99: 419–424. doi:10.1007/s001220051252.

Van Ooijen, J.W., and Voorrips, R.E. 2001. JoinMap® 3.0, software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands.

Yan, H.H., Mudge, J., Kim, D.J., Shoemaker, R.C., Cook, D.R., and Young, N.D. 2004. Comparative physical mapping reveals features of microsynteny between *Glycine max*, *Medicago truncatula*, and *Arabidopsis thaliana*. Genome, 47: 141–155. doi:10.1139/g03-106. PMID:15060611.

Young, N.D., Cannon, S.B., Sato, S., Kim, D.J., Cook, D.R., Town, C.D., et al. 2005. Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. Plant Physiol. 137: 1174–1181. doi:10.1104/pp.104.057034. PMID:15824279.

Zhao, W., Wang, J., He, X., Huang, X., Jiao, Y., Dai, M., et al. 2004. BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. Nucleic Acids Res. 32: D377–D382. doi:10.1093/nar/gkh085. PMID:14681438.